Fall 12-17-2011

# Multivariate Models and Algorithms for Systems Biology

Lipi Rani Acharya

*University of New Orleans*, lacharya@uno.edu

# Multivariate Models and Algorithms for Systems Biology

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Engineering and Applied Science
Computer Science

by

Lipi Rani Acharya

M.Sc., Indian Institute of Technology Madras, 2003
Ph.D., Indian Institute of Technology Kanpur, 2009

December, 2011

# Acknowledgements

I express my deep gratitude to my advisor, Dr. Dongxiao Zhu, for his insightful guidance, endless patience and generous support throughout the tenure of this dissertation. I also thank him for letting me be a part of this exciting project. Numerous discussions and seminars held under his knowledgeable direction, his invaluable expertise and unflinching encouragement have gone a long way in successful completion of this work. It was a great pleasure working with him and I am indebted to him for helping me grow both professionally and personally.

I am grateful to the members in my dissertation committee for their constructive suggestions on this research and their incredible support. Besides Dr. Zhu, I thank Dr. Huimin Chen, Dr. Adlai N DePano, Dr. Linxiong Li and Dr. Christopher M Summa. I also had the privilege of taking classes with them during my course work and what I learned from them had a great impact on this dissertation. I sincerely thank each one of them.

I am thankful to my colleague Thair Judeh with whom I had the opportunity to discuss and collaborate on many research projects. I also acknowledge the initial contributions made by Dr. Zhensheng Duan and the suggestions of Dr. Michael Rabbat in Chapter 5 of this dissertation. I express my thanks to Dr. Guangdi Wang for sharing his biological insights which proved instrumental in testing our methodology in Chapter 6.

This research was supported by the Office of Research and Sponsored Programs at the University of New Orleans and the grant R21LM010137 from National Institute of Health to Dr. Zhu. I acknowledge the support of both the agencies.

My words of appreciation go to Ms. Jeanne Boudreaux, our department secretary, and Ms. Zella Huaracha, the doctoral program coordinator, for their hard work.

I have thoroughly enjoyed the company of a group of wonderful people in Dr. Zhu's

research group, especially Nan Deng, Guorong Xu, Tin Nguyen, Thair Judeh and Kristen M Johnson. They all have taught me many things about life. I appreciate their friendships during the course of this dissertation.

I will always be indebted to my parents, parent-in-laws, brother and sisters for their unconditional love and support, which helped me stay focussed during the years. Finally, I express my most heartfelt thanks to my husband, Sachi Mishra, for his selfless love, for providing me with unending encouragement and for standing by me in all good and bad times.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations and Acronyms

| | |
|---|---|
| ANOVA | Analysis of Variance |
| ARACNE | Algorithm for the Reconstruction of Accurate Cellular Networks |
| BFS | Breadth First Search |
| BIC | Bayesian Information Criterion |
| BNT | Bayes Net Toolbox |
| C3NET | Conservative Causal Core Network |
| CDF | Cumulative Density Function |
| cDNA | Complementary DNA |
| CLR | Context Likelihood of Relatedness |
| DAG | Directed Acyclic Graph |
| DNA | Deoxyribonucleic Acid |
| DPI | Data Processing Inequality |
| DREAM | Dialogue for Reverse Engineering Assessments and Methods |
| E. *coli* | Escherichia *coli* |
| EM | Expectation-Maximization |
| EMBL | The European Molecular Biology Lab |
| ERBB | Avian Erythroblastosis Oncogene B |
| GCRMA | GeneChip RMA |
| GEO | Gene Expression Omnibus |
| GGM | Graphical Gaussian Model |
| GnRH | Gonadotropin-Releasing Hormone |
| GSGA | Gene Set Genetic Algorithm |
| GSGS | Gene Set Gibbs Sampler |

| | |
|---|---|
| GSSA | Gene Set Simulated Annealing |
| HER | Human Epidermal Growth Factor Receptor |
| IF | Information Flow |
| IFGS | Information Flow Gene Set |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| KNN | K-Nearest Neighbors |
| MAS5 | MicroArray Suite 5.0 |
| MBEI | Model-Based Expression Index |
| MCMC | Markov Chain Monte Carlo |
| MI | Mutual Information |
| MIM | Mutual Information Matrix |
| MH | Metropolis-Hastings Algorithm |
| MLE | Maximum Likelihood Estimate |
| mRNA | Messenger RNA |
| MRNET | The Maximum Relevance Minimum Redundance Network |
| MSE | Mean Squared Error |
| NCBI | National Center for Biotechnology Information |
| NEM | Nested Effects Models |
| PBN | Probabilistic Boolean Network |
| PMOM | Progesterone-Mediated Oocyte Maturation |
| PPI | Protein-Protein Interaction |
| PPV | Positive Predictive Value |
| RMA | Robust Multi-Array Average |
| RN | Relevance Network |
| RNA | Ribonucleic Acid |
| SE | Squared Error |
| SGD | The Saccharomyces Genome Database |

# Abstract

Rapid advances in high-throughput data acquisition technologies, such as microarrays and next-generation sequencing, have enabled the scientists to interrogate the expression levels of tens of thousands of genes simultaneously. However, challenges remain in developing effective computational methods for analyzing data generated from such platforms. In this dissertation, we address some of these challenges. We divide our work into two parts. In the first part, we present a suite of multivariate approaches for a reliable discovery of gene clusters, often interpreted as pathway components, from molecular profiling data with replicated measurements. We translate our goal into learning an optimal correlation structure from replicated complete and incomplete measurements. In the second part, we focus on the reconstruction of signal transduction mechanisms in the signaling pathway components. We propose gene set based approaches for inferring the structure of a signaling pathway.

First, we present a constrained multivariate Gaussian model, referred to as the informed-case model, for estimating the correlation structure from replicated and complete molecular profiling data. Informed-case model generalizes previously known blind-case model by accommodating prior knowledge of replication mechanisms. Second, we generalize the blind-case model by designing a two-component mixture model. Our idea is to strike an optimal balance between a fully constrained correlation structure and an unconstrained one. Third, we develop an Expectation-Maximization algorithm to infer the underlying correlation structure from replicated molecular profiling data with missing (incomplete) measurements. We utilize our correlation estimators for clustering real-world replicated complete and incomplete molecular profiling data sets. The above three components constitute the first part of the dissertation. For the structural inference of signaling pathways, we hypothesize a directed signal pathway structure as an ensemble of overlapping and linear signal transduc-

tion events. We then propose two algorithms to reverse engineer the underlying signaling pathway structure using unordered gene sets corresponding to signal transduction events. Throughout we treat gene sets as variables and the associated gene orderings as random. The first algorithm has been developed under the Gibbs sampling framework and the second algorithm utilizes the framework of simulated annealing. Finally, we summarize our findings and discuss possible future directions.

**Keywords** Replicated data, incomplete data, correlation, covariance matrix, multivariate Gaussian mixture models, expectation-maximization (EM) algorithm, gene sets, Gibbs sampling, signaling pathways, signal transduction, discrete optimization, simulated annealing.

# Chapter 1

# Background and Introduction

## 1.1 Molecular Profiling Measurements

Rapid advances in high throughput data acquisition platforms, such as microarrays [46, 82, 128] and next generation sequencing [99, 133], are bringing about a revolution in our understanding of biological complexity. It has become clear that genes do not function alone but through complex biological pathways. Characterization of such intricate pathways can provide deep insights into the biomolecular interaction and regulation mechanisms, which pose several challenges to biology and genetics. Using traditional approaches, which mainly focussed on one gene at a time, it was not feasible to survey the concerted activities of multiple genes simultaneously. Emergence of high throughput technologies have enabled the researchers to interrogate the expression profiles of tens of thousands of genes in a single experiment. An enormous amount of data generated by such platforms can be accessed from public repositories and databases, e.g. National Center for Biological Technology (NCBI) Gene Expression Omnibus (GEO) [15], the European Molecular Biology Lab (EMBL) ArrayExpress [115] and the Saccharomyces Genome Database (SGD) [54]. This has created substantial interest among researchers in the development of effective methodologies for a better understanding of fundamental cell functions and genetic causes of human diseases.

### 1.1.1 Microarray Experiments

Microarrays have become a standard tool for gene expression measurement in the biomedical community. Using microarray chips, it is now possible to capture the genome-wide picture

of an organism under different conditions. Microarrays are useful in a wide range of research areas such as gene screening [120,137], drug discovery [56,63] and pathway analysis [39,112, 113]. Some of the more familiar techniques used in the analysis of microarray data include detection of differentially expressed genes [27, 60, 124], gene clustering [93, 94, 152], sample classification and biomarker discovery [108, 153] and gene network inference [7, 20, 88, 155]. However, the outcome of any of these analyses is directly affected by the quality of gene expression profiles under study. In general, the measurements generated from microarray platforms are contaminated with excessive noise, which may be introduced at various stages of a microarray experiment.

There are a sequence of steps involved in acquiring gene expression profiles using microarray technology, which we briefly describe below:

Chip Manufacturing: A microarray is made of a solid surface on which strands of polynucleotide, also known as probes, are attached or synthesized in fixed locations. Two popular gene expression microarrays are: spotted or cDNA microarrays [128] and oligonucleotide chips (Affymetrix GeneChips) [82]. In cDNA microarrays, probes are mechanically printed on the slide and each probe, which is a cDNA fragment, represents one gene. In the case of Affymetrix chips, probes are directly synthesized on the array. Each probe on a Affymetrix chip is a DNA oligonucleotide. A set of sibling probes, referred to as a probe set, is used to represent one gene.

Sampling and Labeling: A microarray experiment begins with the isolation of RNAs from the subject cells. In cDNA microarrays, RNAs are extracted from both control and experimental samples. RNAs are reverse transcribed into cDNAs. By *in vitro* transcription cDNAs are converted to cRNAs, which are then labeled using fluorescent dyes of two different colors (usually red and green). The labeled transcripts are called targets. Affymetrix microarrays, on the other hand, are single channel platforms which use only one sample per chip.

Hybridization: The basic principles used in microarrays are: (1) DNA and RNA specifically bind to their complementary sequence and (2) the binding occurs in proportion to the abun-

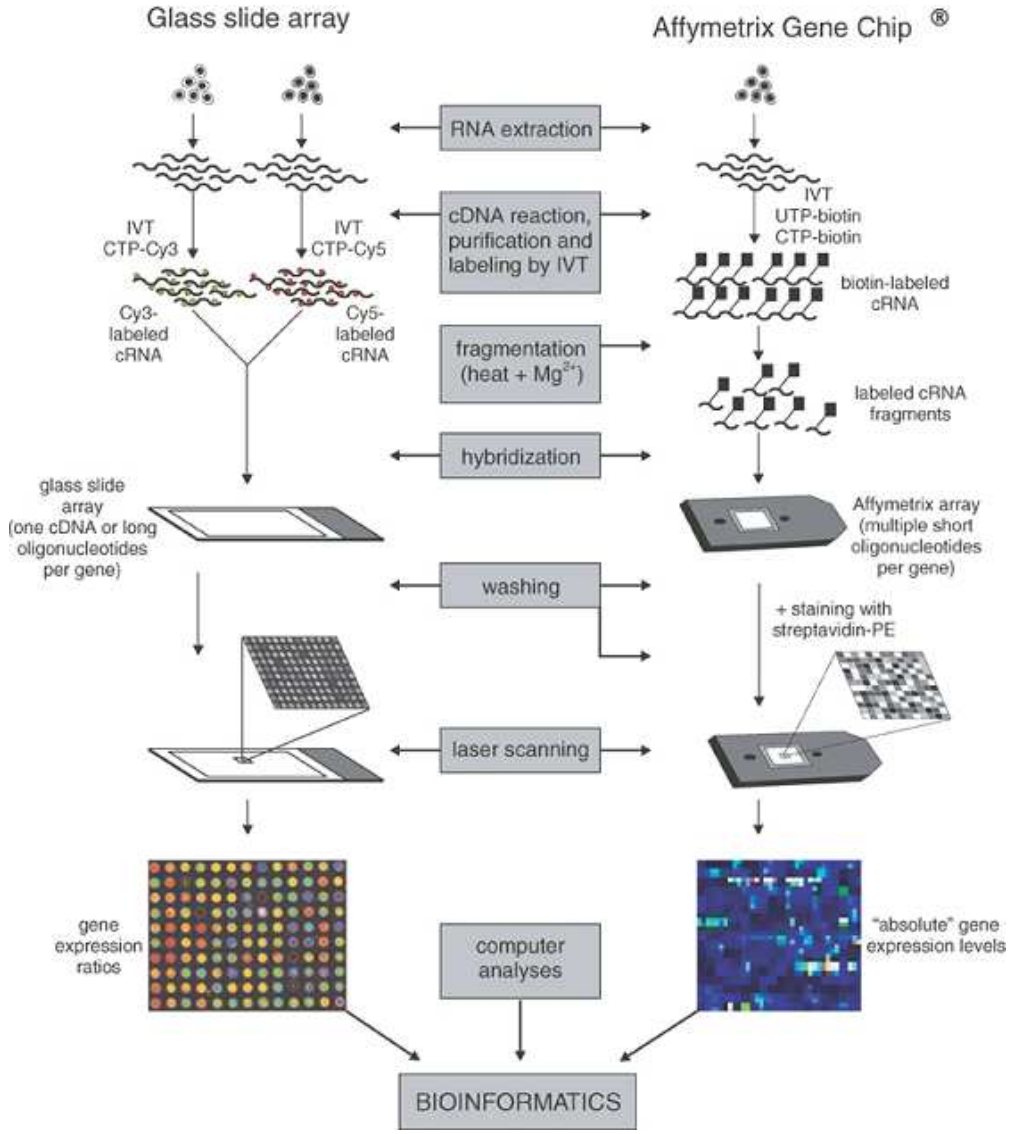


Figure 1.1: Illustration of cDNA (Left) and Affymetrix (Right) microarray technologies. Figure reused by permission from Mcmillan Publishers Ltd: Leukemia [138], copyright 2003.

dance of a sequence. Hybridization is the process by which the labeled targets bind to the probes on the array. After hybridization, microarray is washed to eliminate the portions of unused targets.

Scanning and Imaging: The washed microarray is illuminated using a laser light that causes the labeled targets to emit fluorescence. The emitted fluorescence is scanned and stored as an image which consists of a grid of spots, one for each probe.

Data Acquisition: The image is transformed into numerical values to obtain raw intensities for each probe. Before the raw data can be used for analysis, it is preprocessed using background correction and normalization. Background correction is used to correct the processing effects on the array and make adjustments for cross-hybridization (non-specific bindings). Normalization is used for reducing the within or between array variations. After this step, a matrix comprising of relative or absolute mRNA abundance levels is obtained, which is used for bioinformatics analysis.

Fig. 1.1 demonstrates a step-by-step procedure used in both cDNA and Affymetrix microarray experiments. It is worth mentioning here that the methodologies developed in this dissertation are not restricted to microarray data. They are applicable to other molecular profiling data, such as proteomics data. In the following chapters, however, we have illustrated the performance of our methods using molecular profiling measurements generated from Affymetrix microarray technology. Compared with other microarray platforms, Affymetrix microarrays are often preferred due to a number of reasons:

- Affymetrix microarray is a single-color oligonucleotide array, which results in a simplified experimental design.

- As opposed to using a single long probe, Affymetrix microarrays use a set of short sibling probes for representing a gene. This leads to an increased sensitivity and specificity.

- The probes in a probe set are randomly spread across an Affymetrix microarray. As a

result, the effect of localized artifacts is reduced.

- Affymetrix microarrays have increased throughput and reproducibility.
- A wide range of computational tools are easily available for analyzing Affymetrix microarray data.

### 1.1.2 Replicated Molecular Profiling Data

Replication is commonly used in biomedical experiments to account for the inherent variability and noise in data. The necessity and benefit of replication is more pronounced for high throughput experiments, where data are often exposed to excessive noise. Even in the case of more accurate next-generation deep sequencing data [134], there still exist multiple sources of uncertainty deriving from fragmentation bias, base calling, short-read aligning and short-read counting based on the error-prone genome annotation [99].

The following two types of replications are commonly used in high throughput experiments: *biological replication* and *technical replication* [10]. Biological replication corresponds to the type where measurements from multiple cases are considered, e.g. samples collected from different breast cancer patients. In technical replication, multiple replicates of the same biological replicate are used, e.g. replicated spots representing the same gene on a chip or different aliquots of the same sample used in different chips. Biological replication is useful to measure the variability across population, whereas technical replication is employed for estimating measurement level variability.

The replication mechanism used in underlying experimental design may be either *blind* or *informed* to the data analysts. A good example of the former is the Affymetrix GeneChip [82], where 11 perfect match probes are designed against the 3-prime end of mRNA to interrogate the abundance level of the same gene, although a mixture of gene isoforms can exist. For this reason they are general-sense replicates with blind replication mechanism and large internal variation. A good example of the latter is the Illumina hybridization-based

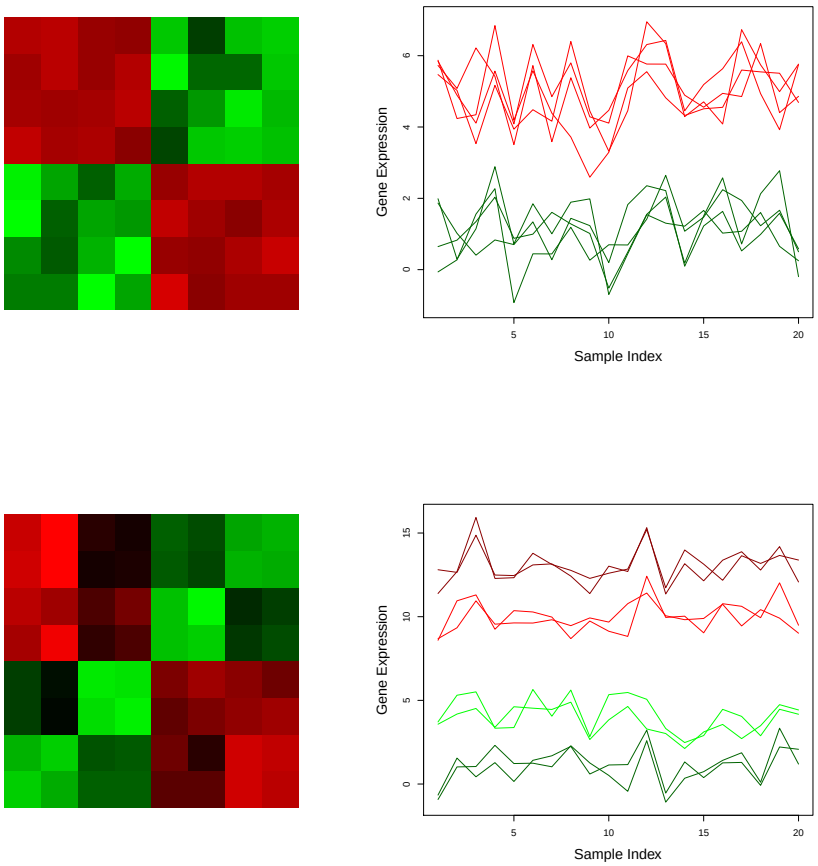


Figure 1.2: Correlation structures (Left) and molecular profiling data (Right) corresponding to a pair of genes, each with 4 replicated measurements. The upper panels represent the correlation structure and molecular profiling data with blind replication mechanism, whereas the lower panels correspond to the ones with informed replication mechanism. In the case of informed replication mechanism 2 biological replicate and 2 technical replicates nested within each biological replicates are used for a gene. Figure reused from [4].

BeadArray [46] and deep sequencing based Genome Analyzer II [134], where $6-12$ samples of whole-genome gene expression are simultaneously profiled for each chip/run. Both biological replicates and/or technical replicates can be used for each chip/run. For this reason they can be treated as narrow-sense replicates with informed replication mechanisms. In many cases, replicates with blind mechanism can be nested within the ones with informed mechanism, and *vice versa* [70].

Both blind and informed replication mechanisms must be considered for a robust pattern analysis of replicated molecular profiling data. For instance, Fig. 1.2 presents two gene sets with the same number of replicated measurements but with different replication mechanisms. As demonstrated in the figure, the corresponding true correlation structures capture the replication mechanisms and are different from each other. In addition to diverse replication mechanisms used in experimental design, high throughput molecular profiling data may also be incomplete, i.e. data may contain a small to large percentage of missing values [81]. Incompleteness may arise due to various reasons such as sample contamination, cross-hybridization, high background noise combined with low signal. In some high throughput molecular profiling experiments such as mass-spectrum, the ratio of missing values can be as high as 30%. Clearly, incomplete replicated measurements present obstacles for further data analysis. It is necessary to design computational frameworks for a reliable pattern discovery from replicated complete and incomplete measurements with both blind and informed replication mechanisms [1, 4, 158, 161].

## 1.2 Pathway Analysis

Molecular profiling measurements generated from microarray experiments are usually in the form of large matrices of gene expression levels measured under different conditions. These measurements only act as a source for investigating biological complexity, they do not themselves reveal the whole picture of this complexity. A follow-up data analysis must be performed to uncover gene interaction and regulation patterns underlying molecular profiling

data. However, gaining biological insights from genome-wide measurements is challenging due to the complexity of biological systems (large number of biomolecules $p$) and availability of an insufficient amount of data that estimate the complex dependency structure (small sample size $N$), a problem referred to as the *curse of dimensionality* [49]. Therefore, an initial characterization of molecular profiling data is required to organize genes into smaller groups on the basis of their expression profiles. Individual gene groups are then analyzed for their potential role in biological pathways.

Pathway-level analysis is the key to make biological inferences and hypothesis from molecular profiling data. A biological pathway represents the biological reactions and biomolecular interaction mechanisms within a cell. In recent years, many annotated biological pathways and tools for their analysis have become increasingly available due to rapid advancements in high-throughput data acquisition methods [30,57,68,140,145]. However, our current knowledge about the signal transduction activities in a cell, which affects gene expressions via downstream transcription factors, is quite limited. For example, the signaling pathway structures available from public databases may not represent a complete picture of underlying signal transduction events among genes that are already known to be related to the pathway. There might exist additional mechanisms among genes present in the pathways. Moreover, the pathways in databases are often generic, whereas scientists are many times interested in learning context-specific signaling pathway structures. We categorize signaling pathway analysis into the following two subproblems: (1) which genes are related to a signaling pathway and (2) how the genes within a pathway interact with each other.

### 1.2.1 Structure of Signaling Pathways

Structural study of signaling pathways is important to improve our understanding of fundamental cell functions, e.g. growth, metabolism, differentiation and apoptosis, which are driven by simultaneous action of several cascades of reactions from the cell surface to the nucleus [6]. According to the central dogma of molecular biology, genetic information is

encoded in double stranded DNA. The information stored in DNA is transferred to single stranded messenger RNA (mRNA) to direct protein synthesis. Signal transduction activities in a pathway are the primary mean to control the passage of biological information from DNA to mRNA with mRNA directing the synthesis of proteins.

A signaling pathway comprises of several overlapping signal transduction events among a set of biomolecules (usually proteins) upstream of transcription factors. Signal transduction events in a pathway are triggered by the binding of external ligands (e.g. cytokine and chemokine) to the transmembrane receptors. This binding results in sequential activations of signaling molecules, such as cytoplasmic protein kinase, to lead to a biological end point (transcription factor). Since activation of signaling pathways affects gene expressions via transcription factors, it is necessary to understand signal transduction mechanisms upstream of transcription factors.

In an abstract sense, the structure of a signaling pathway can be described as a directed graph, where each node represents a protein and a directed edge represents the passage of information from one node to another node. Inference of such directed network topologies is a major challenge in systems biology [24, 122]. Some of the popular pathway databases that comprise of manually curated pathway maps representing our current knowledge on biological networks include KEGG (`www.genome.jp/kegg`), BioCarta (`www.biocarta.com`) and NCBI BioSystems (`www.ncbi.nlm.nih.gov/biosystems`). For a more comprehensive list of web-accessible biological pathway and network databases, we refer to [12].

### 1.2.2 Identification of Signaling Pathway Components

For inferring the structure of a signaling pathway, it is first necessary to identify the set of genes that comprise the pathway. Gene clustering [93, 94, 152] is often one of the first steps used in the identification of pathway components. Gene clustering is a simple data partitioning approach for organizing genes in different groups, where genes within a group share functional similarities. Compared with other similar approaches, results from gene

clustering are often easier to interpret. Another advantage of gene clustering is its applicability in the absence of any prior knowledge about data, such as the functions of individual genes. Some of the popular gene clustering algorithms include: Hierarchial clustering [34], K-means clustering [48] and model-based clustering [90].

A number of other supervised and unsupervised learning algorithms used in the identification of pathway components are differential expression analysis [27, 60, 124], matrix factorization schemes [17, 72], co-expression networking combined with network partitioning [19, 155, 156] and others [52, 142]. Gene lists obtained by an application of any of the aforementioned approaches represent candidate pathway components. The candidate gene lists are statistically tested for their biological significance using over-representation analysis [33, 71] to identify gene ontology terms that are over-represented in the candidate list or functional class scoring [140, 145] which incorporates functional indicators of the genes. This process leads to the discovery of signaling pathway components.

In general, the discovery of biologically meaningful pathway components is highly dependent on the computational approach used in their identification. Since the choice of an approach is often problem specific, a more crucial issue is the use of a reliable metric which can be employed to learn the dependencies among genes in a pathway and can be easily accommodated by diverse pathway learning techniques. Indeed, correlation is one such measure that captures the functional relationships among genes and facilitates the identification of pathway components.

Correlation is at the core of many supervised and unsupervised pattern analyses approaches. In unsupervised learning, many gene clustering algorithms group genes on the basis of their correlation structure [34, 55, 151, 152]. Correlation structure is also employed by many gene networking algorithms to determine the presence or absence of network edges, which is a strong indicator of the functional relevancy between a pair of genes. However, similar to the case of high-dimensional molecular profiling data, it is difficult to draw meaningful conclusions from genome-scale co-expression networks, which may be too broad or

abstract of a representation for a particular biological process of interest. Therefore, learning a finer level of detail from large-scale biological networks is often of more interest to the scientists. As a result, co-expression networking is followed by an application of network clustering [69, 84, 156] or community detection algorithms [5, 73, 78, 105–107, 111]. The resulting subnetworks are interpreted as functional modules or signaling pathway components. In terms of supervised learning, the performance of various model-based classification methods [49], e.g. linear and quadratic discriminate analysis, relies on an accurate estimate of the population correlation structure. These analyses may further be used to learn pathway components and context-specific gene networks in disease groups [39, 103].

## 1.3 Previous Works and Current Challenges

In this dissertation, we develop methodologies to address two major problems in computational systems biology: (1) estimation of an optimal correlation structure which plays a crucial role in the identification of pathway components and (2) reconstruction of signaling pathway structures demonstrating the signal transduction activities in the pathway components.

### 1.3.1 Correlation-Based Discovery of Pathway Components

As discussed earlier, estimation of an optimal correlation structure is essential for a reliable discovery of pathways from molecular profiling data. However, the existing approaches for inferring population correlation structure do not automatically accommodate replicated measurements. Often, a data preprocessing step of averaging over replicated measurements followed by the estimation of bivariate correlation, such as Pearson correlation coefficient, is used [59, 151, 152]. Averaging is not completely satisfactory since it creates a strong bias while reducing the variance among replicates of diverse magnitudes. Averaging may also lead to a significant amount of information loss. For example, useful information including weak patterns and opposite patterns may cancel out by averaging over replicates.

In one-gene-at-a-time analyses for detecting differentially expressed genes between categorical phenotypes (e.g. healthy *versus* cancer tissues), replicates are sufficiently exploited by using Analysis of Variance (ANOVA) type methods, e.g. [70, 147]. However, this type of analysis identifies differentially expressed genes between two phenotypes or experimental conditions without considering the complicated regulatory relationships among genes, which is often reflected in gene-gene correlation structure. Correlation-based analysis, e.g. hierarchical clustering [34, 55], differential correlation [132] and co-expression networking [16, 18, 19, 85, 88, 116, 125, 149, 155, 156], are viable multi-gene approaches to decipher underlying gene regulatory mechanisms and to infer functional modules or pathway components. With few exceptions, the existing clustering and networking algorithms do not explicitly accommodate replicated measurements. Commonly, replicates are averaged (e.g. weighted [59], un-weighted or something in between [151]) or, for Affymetrix data, summarized (e.g. RMA [64], GCRMA [150], MAS5 [58] and Model-Based Expression Index (MBEI) [81]). The averaging and summarizing are necessary since the mean of the replicates is one of the primary interests in one-gene-at-a-time analysis.

In multi-gene clustering and networking analysis, the primary interest is often to estimate a scale-free correlation structure among genes that does not depend on the abundance level of each individual replicate (Fig. 1.3). Expression patterns derived from low abundance profiles can be just as important as those derived from high abundance ones. Averaging or summarizing over replicates of diverse magnitude might wipe out important patterns of low magnitude and/or cancel out patterns of similar magnitude. The averaging or summarizing procedure, originally targeted for differential expression analysis, becomes a nuisance in gene clustering and networking analysis. The situation is even worse when the replication mechanisms used in underlying experimental design is available *a priori* or the replicated measurements contain a small to large percentage of missing values.

With few exceptions, e.g. [93, 94, 158], the existing gene clustering and networking algorithms do not appropriately accommodate replicated and/or incomplete measurements.

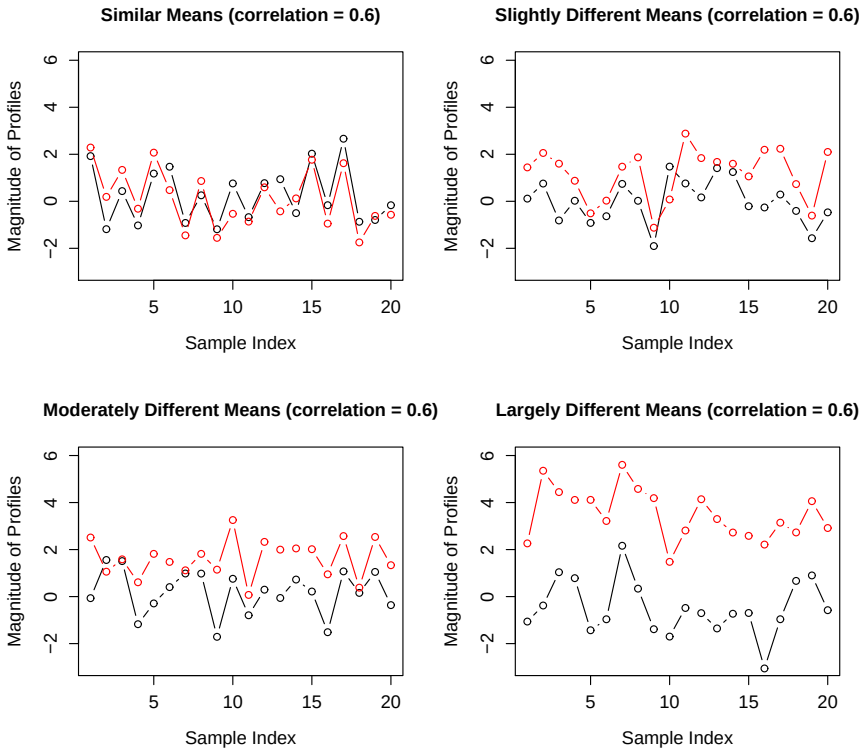


Figure 1.3: Correlation is scale-free. The 4 pairs of non-replicated simulated profiles have the same correlation of 0.6, but differ vastly in their relative magnitude. Figure reused from [161].

The increased power of detecting hidden patterns in data is achieved by sufficiently exploiting the replicates, which has been demonstrated using infinite Bayesian mixture models [93, 94] and parsimonious (or blind-case) multivariate Gaussian models [158]. In the infinite mixture model approach, authors used both an elliptical model that allows within-replicate variation across difference samples to be different and a spherical model that does not. The authors showed that approaches using information about the within-replicate variability (elliptical and spherical models [94]) generally outperform the averaging or summarizing approaches. In the parsimonious multivariate Gaussian models, authors proposed a parsimonious correlation model [158] that shares a similar spirit to the Bayesian elliptical or spherical model approach [94] in that both approaches explicitly consider each replicate, individual variability and their relationships.

Nevertheless, none of the two approaches are ready to analyze replicated and/or incomplete molecular profiling data with a prior known experimental design information. Therefore, it is necessary to design new computational paradigms for an accurate estimation of the correlation structure that allows biomedical researchers to sufficiently exploit replicated complete and incomplete measurements with or without prior knowledge of replication mechanism. This, in turn, is expected to give rise to a reliable discovery of pathway components (Fig. 1.4). **It is one of the two major contributions of this dissertation to address this challenge**.

### 1.3.2 Reconstruction of Signaling Pathway Structures

Reconstruction of signaling pathway structures is essential to decipher complex regulatory relationships in living cells. Characterization of complicated interaction patterns in signaling pathways can provide insights into biomolecular interaction and regulation mechanisms. Consequently, there have been a large body of computational efforts for reconstructing signaling pathway structures using Probabilistic Boolean Networks (PBNs) [135, 136], Bayesian Networks [37, 130], Mutual Information Networks [7, 19, 96], Graphical Gaussian

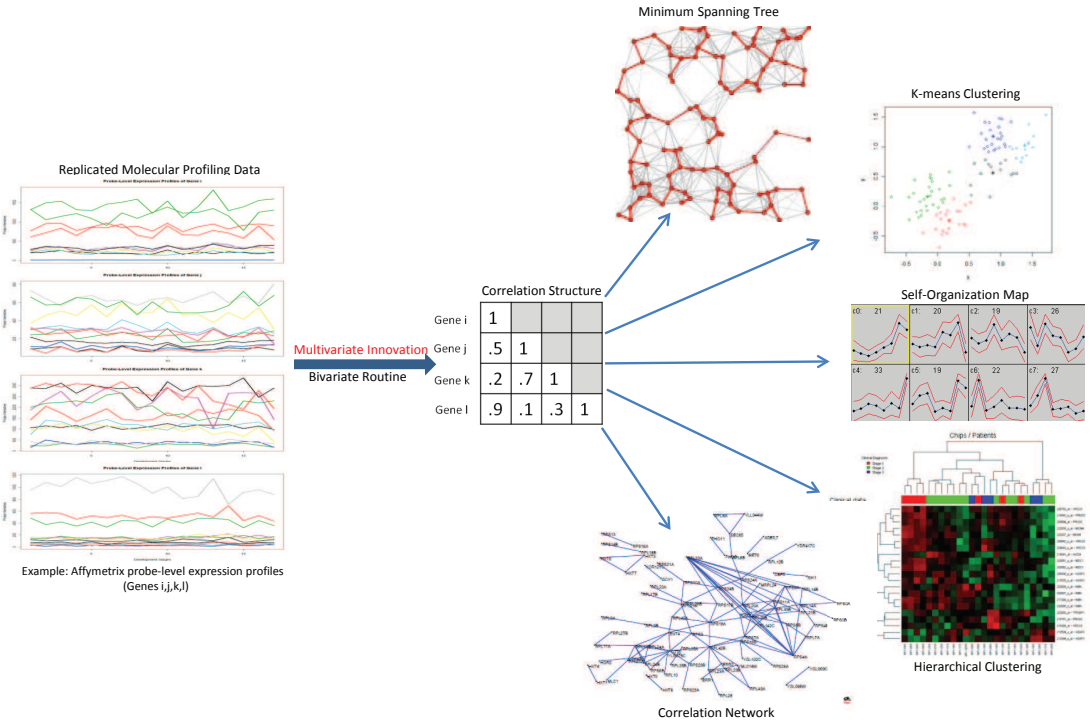


Figure 1.4: Gene clustering and networking using replicated molecular profiling data. The left block represents 4 replicated molecular profiles, in which the magnitude of each molecular profile (one color curve) differs significantly from the others. The middle block displays a scale-free correlation matrix of 4 replicated gene expression profiles. The right side block shows five popular gene clustering and networking algorithms. Figure reused from [161].

Models [32, 75, 125, 126] and other approaches [38, 143, 144, 148, 155].

Although the existing approaches are useful, they often represent a phenomenological graph of the observed data. For example, a parent set of each gene in Bayesian networks indicates statistically causal relationships. In addition, the accuracy of a learned Bayesian network is determined by the choice of the number of parents for each node, a metric used to score a structure and other parameters set to alleviate the non-trivial computational burdens associated with Bayesian network inference. Mutual information networks, graphical Gaussian models and boolean networks are computationally tractable even for large signaling pathways, however the co-expression criteria used in mutual information networks and graphical Gaussian models only models a possible functional relevancy, whereas the use of boolean functions in boolean networks may lead to an oversimplification of the underlying gene regulatory mechanisms. Moreover, the aforementioned approaches do not explicitly consider signal transduction events characterizing a signaling pathway. Signal transduction events refer to directed linear cascades of reactions from the cell surface to the nucleus and form the basic building blocks of a signaling pathway. It is, therefore, necessary to design computational approaches for the structural inference of signaling pathways by incorporating signal transduction mechanisms.

With few exceptions in the field of communication networks, the existing network inference approaches do not explicitly accommodate signal transduction events. The frequency method in [118] assumes a tree structure in the paths between pairs of nodes (genes). However, the method is subjected to fail in the presence of multiple paths between the same pair of nodes. The cGraph algorithm presented in [76] adds weighted edges between each pair of nodes that appear in some set of gene co-occurrence and so the networks inferred by this approach might contain a large number of false positives. The EM approach [119, 157] treats permutations of genes in a signal transduction as missing data and infers a network by assuming a linear arrangement of genes along with a prior knowledge of two end nodes. It is also difficult to incorporate prior knowledge about regulator-target pairs in the approaches

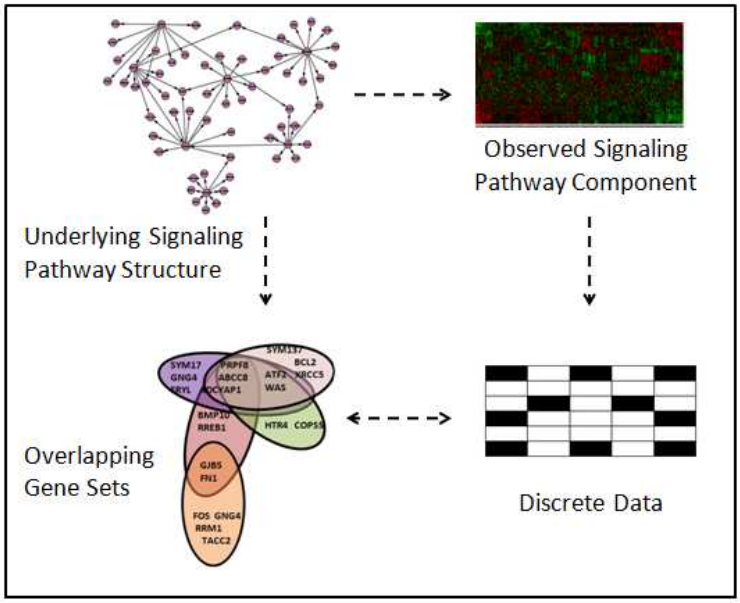


Figure 1.5: Representation of a gene set compendium as binary discrete data and *vice versa*.

mentioned above.

A central aspect of developing such network reconstruction approaches is to understand the structure of signaling pathways, which are an ensemble of several overlapping signal transduction events with a linear arrangement of genes in each event. Overlapping arises from simultaneous participation of genes in many biological processes. In the present context, we refer to the set of genes in a signal transduction event as an *information flow gene set* (IFGS). Thus, an IFGS only reflects which genes participate in a signal transduction event, but not their ordering. The underlying signaling pathway structure can be reconstructed by inferring the order of genes in each IFGS.

An IFGS can also be interpreted as a discrete set of genes expressed in an experiment, whereas an IFGS compendium comprises of many overlapping gene sets corresponding to different experiments. With this understanding, an IFGS based approach can be compared with other network inference algorithms which accommodate discrete measurements, such as Bayesian networks [26, 37, 130] and mutual information networks [7–9, 19]. Fig. 1.5 sketches the equivalence between an IFGS compendium and binary discrete molecular profiling measurements obtained by considering the presence (expressed) or absence (not

expressed) of genes in a gene set. However, an IFGS approach may be more suitable for capturing higher order signal transduction mechanisms as opposed to pairwise interactions or causal interactions. Compared with other network inference approaches which utilize continuous molecular profiling data, an IFGS approach may be more robust to noise and may facilitate data integration from multiple data acquisition platforms.

It is also worth mentioning the difference between the concept of an IFGS presented here and a *gene signature* used in the literature. A gene signature usually corresponds to a set of genes with combined pattern of expression downstream of transcription factors and is often linked to a given biological state of interest. An IFGS, on the other hand, represents a set of molecules (usually proteins) in a signaling pathway upstream of transcription factors which participate in a signal transduction event in the pathway. Moreover, IFGSs related to a signaling pathway indicate the existence of an underlying structure, whereas a gene signature may only correspond to a set of functionally relevant genes without suggesting the presence of a structure. Gene signature based analysis has received much attention in recent years. The relative advantages of working with gene signatures in bioinformatics analyses have been adequately demonstrated [112, 113, 121, 140]. They have also been used to dissect drug mechanism of action and to find transcriptional connections among genes, drugs and diseases [63, 77]. However, signaling pathway structure inference by sufficiently exploiting gene sets corresponding to signal transduction mechanisms, a promising area of bioinformatics research, remains underdeveloped. **It is the second major contribution of this dissertation to address this challenge**.

## 1.4 Outline of Dissertation

The goal of this dissertation is to develop novel methodologies for inferring gene association and regulation patterns from molecular profiling data. The work presented here is composed of two parts. The first part presents a sequence of multivariate approaches leading to a reliable discovery of gene clusters, often interpreted as pathway components, from replicated

molecular profiling data. Our approach is to learn an optimal correlation structure from both replicated complete and incomplete molecular profiling data (Fig. 1.4). In the second part, we address the problem of inferring the structure underlying a signaling pathway component. We develop algorithms by treating gene sets corresponding to signal transduction activities in a signaling pathway as the basic building blocks of the underlying structure (Fig. 1.5).

In Fig. 1.6, we sketch a flowchart of the problems considered in this dissertation. In Chapters 2 and 3, we develop two generalized multivariate correlation estimators for pattern discovery from replicated and complete molecular profiling data. In Chapter 2, specifically, we present a correlation estimator by explicitly taking into account the prior knowledge of replication mechanisms. We further generalize this correlation estimator in Chapter 3 by designing a finite mixture model. Chapter 4 deals with the problem of inferring correlation structure from replicated and incomplete molecular profiling data. We consider replicated and incomplete measurements with either blind or informed replication mechanisms and develop an EM algorithm to estimate the correlation structure. In Chapters 5 and 6, we present two gene set based algorithms to infer underlying signaling pathway structure in a given pathway component. Chapter 5 presents a sampling based approach by employing the Gibbs sampling framework, whereas Chapter 6 presents a search strategy under the framework of simulated annealing. Finally, in Chapter 7, we summarize our findings and discuss future works.

## 1.5 List of Publications

*Peer-reviewed Journal Papers*

Zhu D, **Acharya L** and Zhang H. A generalized multivariate approach to pattern discovery from replicated and incomplete genome-wide measurements, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(5):1153-1169, 2011.

**Acharya L**, Judeh T, Duan Z, Rabbat M and Zhu D. GSGS: A computational framework to

reconstruct signaling pathway structures from gene sets, Accepted to appear in *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, Preprint `arXiv:1101.3983v3`.

**Acharya L**, Judeh T, Wang G and Zhu D. Optimal structural inference of signaling pathways from overlapping and unordered gene sets, Submitted to *Bioinformatics*, Revised Aug 2011.

*Book Chapters*

**Acharya L** and Zhu D. Multivariate models and algorithms for learning correlation structures from replicated molecular profiling data, In a chapter of the book *Advanced Biomedical Engineering*, InTech 2011.

**Acharya L**, Judeh T and Zhu D. A survey of computational approaches to biological network reconstruction and partition, To appear in a chapter of the book *Machine Learning Approach for Network Analysis: Novel Graph Classes for Classification Techniques*, Wiley Publisher 2011.

*Peer-reviewed Conference Papers*

**Acharya L**, Judeh T, Duan Z and Zhu D. A novel computational framework to reconstruct gene regulatory networks, In the proceedings of *Biotechnology and Bioinformatics Symposium (BIOT)*, 80-81, 2010.

Judeh T, **Acharya L** and Zhu D. Gene network inference via linear path augmentation, In the proceedings of *Biotechnology and Bioinformatics Symposium (BIOT)*, 46-48, 2010.

**Acharya L** and Zhu D. Estimating an optimal correlation structure from replicated molecular profiling data using finite mixture models, In the proceedings of *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 119-124, 2009.

Zhu D, Xu G and **Acharya L**. A generalized multivariate approach for correlation-based pattern discovery from replicated molecular profiling data, In the proceedings of *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 405-410, 2009.

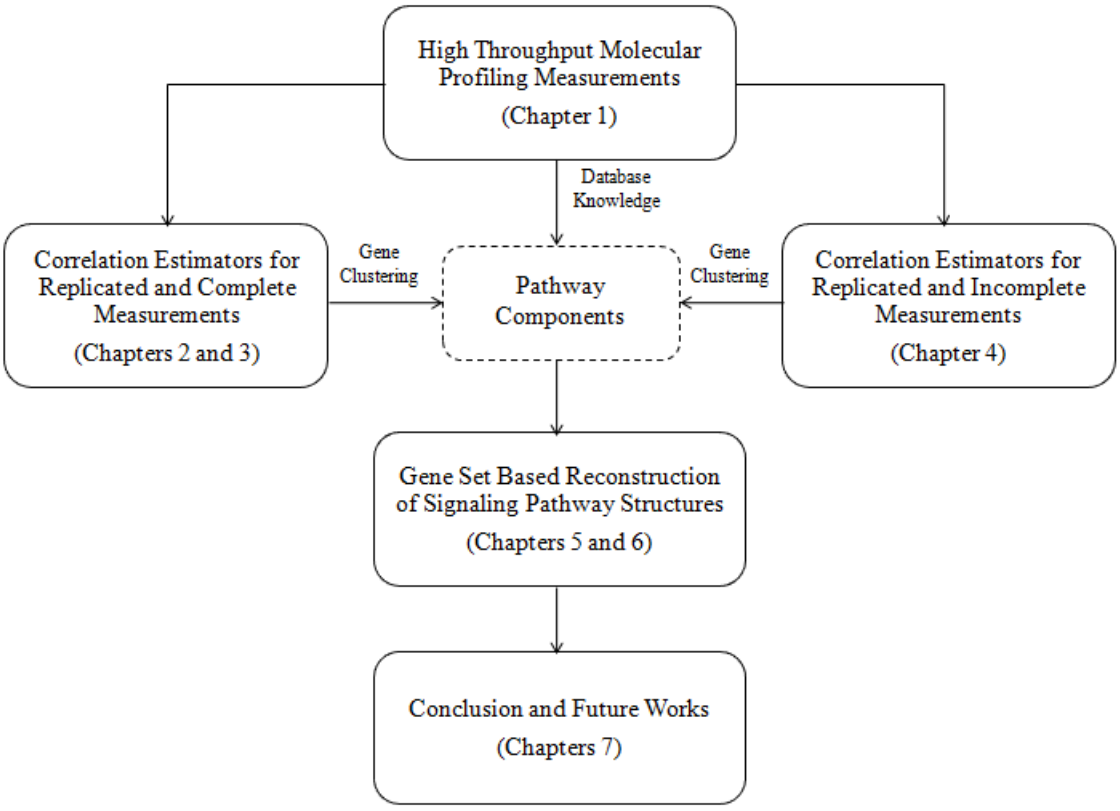


Figure 1.6: An outline of dissertation.

# Chapter 2

# Learning Correlation Structures from Replicated and Complete Molecular Profiling Data I

## 2.1 Introduction

Estimation of an optimal correlation structure is crucial in the pattern analyses of replicated molecular profiling data. Many of these analyses facilitate the identification of pathway components, such as gene clustering [34, 93, 94, 151, 152], co-expression networking [20, 125, 155] combined with network partitioning [5, 69, 105–107, 156] to discover network modules. With few exceptions [93, 94, 158], however, the current approaches for estimating the correlation structure from molecular profiling data do not automatically accommodate replicated measurements. Often, a preprocessing step of summarizing or averaging over replicated measurements is used to reduce the multivariate structure of replicated data into a bivariate one [151, 152]. Bivariate data fits into the framework of pairwise correlation analysis, e.g. Pearson's correlation, which is simple in approach and is achievable at a very low computational cost. However, summarizing or averaging may lead to a significant amount of information loss due to diverse magnitudes of replicated measurements. The situation is worse when the experimental design that explains the replication mechanism of molecular profiling data is known *a priori* but this information is not exploited in the pattern analysis. Therefore, it is necessary to develop computational methods by exploiting each replicate individually and utilizing the prior knowledge of replication mechanisms.

In this chapter[1], we present a generalized multivariate model for estimating the correlation structure of a gene set with replicated and complete molecular profiling measurements. The proposed model, referred to as the informed-case model, generalizes previously known parsimonious or blind-case correlation estimator [1,4,158] by accommodating prior knowledge of replication mechanisms. In many cases, prior knowledge of the number of biological and technical replicates used in underlying experimental design may be known. A straightforward application of blind-case model, which does not distinguish between biological replicates of a gene, may not exploit all underlying information within data. Informed-case model, on the other hand, explicitly incorporates a prior known replication mechanisms in its setting and considers different parameters for different biological replicates of a gene.

Throughout this chapter, we follow and extend the path of blind-case approach. The main reasons are as follows: (1). blind-case model presents a parsimonious multivariate correlation estimator and a closed-form formula for each pair of genes, which successfully alleviates the computational burden derived from dimensionality. Other approaches, such as infinite Bayesian mixture models [93, 94], often represent a computationally daunting task, especially for high dimensional data with an increased number of genes and replicates. (2). blind-case model uses a scale-free correlation structure to separate the estimate of correlation between replicates (the primary interest) from the estimate of within-replicate variability (nuisance). These advantages make the blind-case approach more suitable than infinite Bayesian mixture models for analyzing replicated measurements with large within-replicate variability. However, neither of them are ready to analyze replicated molecular profiling data with a prior known experimental design.

## 2.2 Notations

Throughout this chapter, $G_1, \ldots, G_k$ denote arbitrary biomolecules with $g_{ij}^l$ as their respective abundance levels in the $i^{th}$ replicate (row) and $j^{th}$ sample (column), for $l = 1, \ldots, k$,

[1]Published work [161]. Reused with permission. Copyright, IEEE. All rights reserved.

where the abundance levels are measured over $n$ independent samples. Further, we assume that $m_l$ replicated measurements are available for $G_l$ in each sample, $l = 1, \ldots, k$. The $j^{th}$ column of the given replicated data set is written as $Z_j = (Z_{j1}, \ldots, Z_{jk})^T = (g^1_{j1}, \ldots, g^1_{jm_1}, \ldots, g^k_{j1}, \ldots, g^k_{jm_k})^T$, $j = 1, 2, \ldots, n$, and is assumed to be an independently and identically distributed sample from a multivariate normal distribution with $\sum_{l=1}^{k} m_l$ random variables.

## 2.3 The Existing Blind-Case Approach

### 2.3.1 The Model

We first review the existing blind-case model for estimating the correlation structure of a gene set with replicated and complete measurements. Blind-case model was introduced in [158] and further studied in [1,4]. In this model, each replicate is treated individually as a random variable and data are assumed as random samples from a multivariate Gaussian distribution, which we denote by $N(\mu^B, \Sigma^B)$. We designate the model as 'blind' since it imposes a fixed number of correlation parameters in the underlying correlation structure. The parameters for the blind-case model are defined as

$$\mu^B = \begin{bmatrix} \mu^B_{g_1} e_{m_1} \\ \vdots \\ \mu^B_{g_k} e_{m_k} \end{bmatrix} \tag{2.1}$$

where $\mu^B_{g_i}$ is a scalar and $e_{m_i} = (1, \ldots, 1)^T$ is a vector of size $m_i \times 1$, for $i = 1, \ldots, k$. The correlation matrix $\Sigma^B$ is of size $\sum_{i=1}^{k} m_i \times \sum_{i=1}^{k} m_i$ and is given by

$$\Sigma^B = \begin{bmatrix} 1 & \dots & \rho_{11} & \dots & \rho_{1k} & \dots & \rho_{1k} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{11} & \dots & 1 & \dots & \rho_{1k} & \dots & \rho_{1k} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{k1} & \dots & \rho_{k1} & \dots & 1 & \dots & \rho_{kk} \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \rho_{k1} & \dots & \rho_{k1} & \dots & \rho_{kk} & \dots & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \Sigma^B_{g_1g_1} & \dots & \Sigma^B_{g_1g_k} \\ \vdots & \vdots & \vdots \\ \Sigma^B_{g_1g_k}{}^T & \dots & \Sigma^B_{g_kg_k} \end{bmatrix}. \tag{2.2}$$

In Eq. 2.2, $\Sigma^B_{g_ig_j}$ represents a submatrix of size $m_i \times m_j$, which is defined in terms of a single correlation parameter $\rho_{ij}$. When $i = j$, the parameters $\rho_{ij}$'s correspond to within-molecular correlation. Otherwise, they represent between-molecular correlations. Due to the symmetric nature of a correlation matrix, we assume $\rho_{ij} = \rho_{ji}$. From a practical point of view, between-molecular correlations are more important. Within-molecular correlations, on the other hand, are indicative of data quality. Higher values of within-molecular correlations represent cleaner data.

### 2.3.2 Parameter Estimation

For estimating model parameters from replicated measurements, blind-case approach follows the path of maximum likelihood estimation. Maximum likelihood estimates (MLEs) are frequently used in parameter estimation problems when the underlying distribution is multivariate normal [21]. In such cases, MLEs often have some optimal properties. For example, the MLEs of mean vector and correlation matrix become asymptotically efficient.

When $n > \sum_{i=1}^{k} m_i$, the likelihood function can be written as

$$L(\mu^B, \Sigma^B) = \prod_{j=1}^{n} N(Z_j|\mu^B, \Sigma^B) = \frac{1}{(2\pi)^{\frac{1}{2}(\sum_{i=1}^{k} m_i)n}|\Sigma^B|^{\frac{1}{2}n}} \; e^{[-\frac{1}{2}\sum_{j=1}^{n}(Z_j-\mu^B)^T\Sigma^{B\,-1}(Z_j-\mu^B)]}. \tag{2.3}$$

The MLE's are obtained by maximizing $L$ with respect to $\mu^B$ and $\Sigma^B$. This is achieved by solving

$$d\mathcal{L}/d\mu^B_{g_l} = 0, \tag{2.4}$$

for $l = 1, \ldots, k$ and

$$d\mathcal{L}/d\Sigma^B = 0, \tag{2.5}$$

where $\mathcal{L} = \log L$. This leads to

$$\hat{\mu}^B_{g_l} = \frac{1}{n}\frac{1}{m_l}\sum_{j=1}^{n}\sum_{i=1}^{m_l} g^l_{ij} \tag{2.6}$$

for $l = 1, \ldots, k$. Thus, the MLE of $\mu^B$ is

$$\hat{\mu}^B = \begin{bmatrix} \hat{\mu}^B_{g_1} e_{m_1} \\ \vdots \\ \hat{\mu}^B_{g_k} e_{m_k} \end{bmatrix}. \tag{2.7}$$

The MLE of $\Sigma^B$ is given by

$$\hat{\Sigma}^B = \frac{1}{n}\sum_{j=1}^{n}(Z_j - \hat{\mu}^B)(Z_j - \hat{\mu}^B)^T. \tag{2.8}$$

Since the parameters $\hat{\rho}_{ij}$'s may not be tractable in practice, they are estimated using

$$\hat{\rho}_{ij} = Avg(\hat{\Sigma}^B_{ij}),\ i, j = 1, \ldots, k. \tag{2.9}$$

Eqs. 2.6-2.9 give the MLEs of parameters for the blind-case model. In the case of two

biomolecules, the representation of blind-case model in Eq. 2.1 and Eq. 2.2 coincides with the one presented in [158], which is defined in terms of two within-molecular and one between molecular correlation parameters.

Blind-case model is simple in approach and is especially useful in the case of replicated measurements with no prior knowledge of underlying replication mechanisms. Since the MLEs of model parameters are represented in closed-forms, blind-case approach also benefits from a much reduced computational load than other multivariate approaches, such as infinite Bayesian mixture models [93, 94]. However, blind-case model suffers from two major limitations. First, it may simplify the correlation structure of a gene set with many pairwise gene correlation structures (Eq. 2.2) and second, the model may be overly constrained for some replicated data, e.g. data with a prior known replication mechanism. To overcome these issues, it is necessary to consider a more relaxed multivariate model.

## 2.4 Informed-Case Approach

### 2.4.1 The Model

In this section, we design a multivariate model for estimating the correlation structure from replicated and complete molecular profiling measurements corresponding to a gene set with a prior known replication mechanism. For simplicity, we illustrate the model for the case of 2 genes $G_1$ and $G_2$, where 3 biological replicates with 2 technical replicates nested within each biological replicate are available for both $G_1$ and $G_2$. Throughout we assume that data are independently and identically distributed samples from a multivariate normal distribution $N(\mu^I, \Sigma^I)$. In the above specified case, $Z_j$'s follow a 12-variate normal distribution with

$$\mu^I = \left(\mu^1_{g_1}, \mu^1_{g_1}, \mu^2_{g_1}, \mu^2_{g_1}, \mu^3_{g_1}, \mu^3_{g_1}, \mu^1_{g_2}, \mu^1_{g_2}, \mu^2_{g_2}, \mu^2_{g_2}, \mu^3_{g_2}, \mu^3_{g_2}\right)^T, \tag{2.10}$$

a $12 \times 1$ vector defined in terms of 6 parameters. The correlation matrix $\Sigma^I$ is of size $12 \times 12$ with 16 parameters and is given by

$$\Sigma^I = \begin{pmatrix}
1 & \rho^{tt} & \rho_{g_1}^{12} & \rho_{g_1}^{12} & \rho_{g_1}^{13} & \rho_{g_1}^{13} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{13} \\
\rho^{tt} & 1 & \rho_{g_1}^{12} & \rho_{g_1}^{12} & \rho_{g_1}^{13} & \rho_{g_1}^{13} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{13} \\
\rho_{g_1}^{21} & \rho_{g_1}^{21} & 1 & \rho^{tt} & \rho_{g_1}^{23} & \rho_{g_1}^{23} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{23} \\
\rho_{g_1}^{21} & \rho_{g_1}^{21} & \rho^{tt} & 1 & \rho_{g_1}^{23} & \rho_{g_1}^{23} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{23} \\
\rho_{g_1}^{31} & \rho_{g_1}^{31} & \rho_{g_1}^{32} & \rho_{g_1}^{32} & 1 & \rho^{tt} & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{33} & \rho_{g_1g_2}^{33} \\
\rho_{g_1}^{31} & \rho_{g_1}^{31} & \rho_{g_1}^{32} & \rho_{g_1}^{32} & \rho^{tt} & 1 & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{33} & \rho_{g_1g_2}^{33} \\
\rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{31} & 1 & \rho^{tt} & \rho_{g_2}^{12} & \rho_{g_2}^{12} & \rho_{g_2}^{13} & \rho_{g_2}^{12} \\
\rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{11} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{21} & \rho_{g_1g_2}^{31} & \rho_{g_1g_2}^{31} & \rho^{tt} & 1 & \rho_{g_2}^{12} & \rho_{g_2}^{12} & \rho_{g_2}^{13} & \rho_{g_2}^{12} \\
\rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{32} & \rho_{g_2}^{21} & \rho_{g_2}^{21} & 1 & \rho^{tt} & \rho_{g_2}^{23} & \rho_{g_2}^{23} \\
\rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{12} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{22} & \rho_{g_1g_2}^{32} & \rho_{g_1g_2}^{32} & \rho_{g_2}^{21} & \rho_{g_2}^{21} & \rho^{tt} & 1 & \rho_{g_2}^{23} & \rho_{g_2}^{23} \\
\rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{33} & \rho_{g_1g_2}^{33} & \rho_{g_2}^{31} & \rho_{g_2}^{31} & \rho_{g_2}^{32} & \rho_{g_2}^{32} & 1 & \rho^{tt} \\
\rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{13} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{23} & \rho_{g_1g_2}^{33} & \rho_{g_1g_2}^{33} & \rho_{g_2}^{31} & \rho_{g_2}^{31} & \rho_{g_2}^{32} & \rho_{g_2}^{32} & \rho^{tt} & 1
\end{pmatrix}, \tag{2.11}$$

where the parameters $\rho_{g_1g_2}^{ij}$ and $\rho_{g_1}^{ij}$ (or $\rho_{g_2}^{ij}$) represent intermolecular and intramolecular correlation between the $i^{th}$ and $j^{th}$ biological replicates, respectively. In general, the technical replicates of a biological replicate are highly correlated. Therefore, we use a single parameter $\rho^{tt}$ to represent the correlation between the technical replicates of a biological replicate. We keep this parameter same across all biological replicates of the two genes. However, the model can be made more flexible by assuming this parameter to be different for different biological replicates. It is easy to see that $\Sigma^I$ in Eq. 2.11 is composed of several $2 \times 2$ matrices, each of which are defined in terms of a single correlation parameter. We denote these blocks by $\Sigma_{uv}^{rs}$ for $u,\ v \in \{g_1, g_2\},\ r,\ s \in \{1, 2, 3\}$, where $\rho_{g_1g_1}^{ij} = \rho_{g_1}^{ij}$, $\Sigma_{g_1g_1}^{ij} = \Sigma_{g_1}^{ij}$ and so on. The representation in Eq. 2.11 can be naturally extended to the case of a gene set with a given number of biological replicates and nested technical replicates.

### 2.4.2 Parameter Estimation

Let us consider the case of $J_{m_1}$ and $J_{m_2}$ biological replicates for the genes $G_1$ and $G_2$, respectively. Further, we assume that the number of technical replicates nested within the $j_{m_1}^{th}$ biological replicate is $I_{m_1}^{j},\ 1 \leq j_{m_1} \leq J_{m_1}$ and within $j_{m_2}^{th}$ biological replicate is $I_{m_2}^{j},\ 1 \leq j_{m_2} \leq J_{m_2}$, where

$$\sum_{j=1}^{J_{m_1}} I_{m_1}^{j} = m_1 \ \text{ and } \ \sum_{j=1}^{J_{m_2}} I_{m_2}^{j} = m_2. \tag{2.12}$$

As in the case of blind-case model, we derive the MLEs of $\mu^I$ and $\Sigma^I$ by maximizing $L(\mu^I, \Sigma^I)$. The MLEs are given by (see Appendix A.1 for mathematical proofs)

$$\hat{\mu}_{g_1}^{j_{m_1}} = \frac{1}{I_{m_1}^{j} n} \sum_{k=1}^{n} \sum_{i=\sum_{l=1}^{j} I_{m_1}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_1}^{l}} g_{ik}^{1},\ 1 \leq j_{m_1} \leq J_{m_1} \tag{2.13}$$

and

$$\hat{\mu}_{g_2}^{j_{m_2}} = \frac{1}{I_{m_2}^{j} n} \sum_{k=1}^{n} \sum_{i=\sum_{l=1}^{j} I_{m_2}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_2}^{l}} g_{ik}^{2},\ 1 \leq j_{m_2} \leq J_{m_2} \tag{2.14}$$

This leads to

$$\hat{\mu}^{I^{[1]}} \qquad\qquad\qquad\qquad \hat{\mu}^{I^{[2]}}$$

$$\hat{\mu}^I = \Big( \overbrace{\underbrace{\hat{\mu}_{g_1}^1, \ldots, \hat{\mu}_{g_1}^1}_{I_{m_1}^1 \ times}, \ldots, \underbrace{\hat{\mu}_{g_1}^{J_{m_1}}, \ldots, \hat{\mu}_{g_1}^{J_{m_1}}}_{I_{m_1}^{J_{m_1}} \ times}}^{}, \overbrace{\underbrace{\hat{\mu}_{g_2}^1, \ldots, \hat{\mu}_{g_2}^1}_{I_{m_2}^1 \ times}, \ldots, \underbrace{\hat{\mu}_{g_2}^{J_{m_2}}, \ldots, \hat{\mu}_{g_2}^{J_{m_2}}}_{I_{m_2}^{J_{m_2}} \ times}}^{} \Big)^T \tag{2.15}$$

The MLE of $\Sigma^I$ is

$$\hat{\Sigma}^I = \frac{1}{n} \sum_{j=1}^{n} \begin{bmatrix} (Z_j^{[1]} - \hat{\mu}^{I^{[1]}})(Z_j^{[1]} - \hat{\mu}^{I^{[1]}})^T & (Z_j^{[1]} - \hat{\mu}^{I^{[1]}})(Z_j^{[2]} - \hat{\mu}^{I^{[2]}})^T \\ (Z_j^{[2]} - \hat{\mu}^{I^{[2]}})(Z_j^{[1]} - \hat{\mu}^{I^{[1]}})^T & (Z_j^{[2]} - \hat{\mu}^{I^{[2]}})(Z_j^{[2]} - \hat{\mu}^{I^{[2]}})^T \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\Sigma}^I_{g_1} & \hat{\Sigma}^I_{g_1g_2} \\ \hat{\Sigma}^{I^T}_{g_1g_2} & \hat{\Sigma}^I_{g_2} \end{bmatrix} \tag{2.16}$$

where $Z_j^{[1]}$ and $Z_j^{[2]}$ represent the parts of $Z_j$ containing the measurements for $G_1$ and $G_2$, respectively. In Eq. 2.16, the block $\hat{\Sigma}^I_{g_1g_2}$ presents intermolecular correlations between different biological replicates of the genes $G_1$ and $G_2$. It comprises of the sub-blocks $\hat{\Sigma}^{ij}_{g_1g_2}$, each of which is defined in terms of $\hat{\rho}^{ij}_{g_1g_2}$. The value of $\hat{\rho}^{ij}_{g_1g_2}$ is obtained by averaging the entries in $\hat{\Sigma}^{ij}_{g_1g_2}$.

### 2.4.3 Model Summarization

If we use the method of averaging, as done in Eq. 2.9, to obtain an estimate of $\hat{\rho}$ from the block $\hat{\Sigma}^I_{g_1g_2}$ in Eq. 2.16, the correlation estimate obtained from the informed-case model coincides with the one from blind-case model (see Appendix A.2 for a mathematical proof). Thus, the method of averaging $\hat{\rho} = Avg(\hat{\Sigma}^I_{g_1g_2})$ undermines the experimental design information. It is, therefore, necessary to consider methods other than averaging for accessing the level of pairwise correlation. We adapt two likelihood ratio test based methods [11] to distinguish between the performance of the blind-case and informed-case models.

#### Method 1

We first test the hypothesis that intermolecular correlation between $G_1$ and $G_2$ vanishes. This means, for the parameters $\mu$ ($\mu^I$ or $\mu^B$) and $\Sigma$ ($\Sigma^I$ or $\Sigma^B$), we test the hypotheses

$$H_0 : Z \in N(\mu, \Sigma_0) \quad \text{against} \quad H_\alpha : Z \in N(\mu, \Sigma),$$

where $\Sigma$ and $\Sigma_0$ have the following forms

$$\Sigma = \begin{bmatrix} \Sigma_{g_1} & \Sigma_{g_1g_2} \\ \Sigma^T_{g_1g_2} & \Sigma_{g_2} \end{bmatrix} \quad \text{and} \quad \Sigma_0 = \begin{bmatrix} \Sigma_{g_1} & 0 \\ 0 & \Sigma_{g_2} \end{bmatrix}. \tag{2.17}$$

The summarization statistic for testing the correlation structures of $\Sigma_0$ and $\Sigma$ is given by

$$\Psi = -2 \ \log(\wedge) \tag{2.18}$$

where

$$\wedge = \frac{|\hat{\Sigma}_0|^{-n/2} \exp(\frac{-1}{2} \sum_{j=1}^{n} (Z_j - \hat{\mu})^T \hat{\Sigma}_0^{-1} (Z_j - \hat{\mu}))}{|\hat{\Sigma}|^{-n/2} \exp(\frac{-1}{2} \sum_{j=1}^{n} (Z_j - \hat{\mu})^T \hat{\Sigma}^{-1} (Z_j - \hat{\mu}))}. \tag{2.19}$$

Using Eq. 2.19, we obtain the summarization statistics $\Psi^B = -2 \log \wedge^B$ and $\Psi^I = -2 \log \wedge^I$ for the blind-case and informed-case models, respectively. Under null hypothesis, the two statistics follow an asymptomatic chi-square distribution with 1 and $J_{m_1} J_{m_2}$ degrees of freedom, respectively [11]. We compare the performance of blind-case and informed-case model in terms of $\Phi = -\log_2 P$, where $P$ stands for $P$-value calculated from chi-square distribution. A lower value of $P$ or a higher value of $\Phi$ is indicative of better model performance.

**Method 2**

In this method, we calculate the difference, $\Psi^I - \Psi^B$, of the two likelihood ratio test statistics and compare it to a chi-square distribution with $J_{m_1} J_{m_2} - 1$ degrees of freedom. Small values of $P$, e.g. $P < 0.05$, indicate a better performance of the informed-case model compared with the blind-case model.

It is clear from the above discussions that the correlation structure for the informed-case model (Eq. 2.11) is a generalization of the structure for the blind-case model (Eq. 2.2). In the case of unknown replication mechanisms, informed-case correlation estimator reduces to blind-case estimator.

## 2.5 Results

### 2.5.1 Parameter Settings

We use $\Phi$ to compare the performance of the blind-case and informed-case correlation estimators. As molecular profiling measurements vary in terms of the number of replicates,

sample size and data quality, we consider different combinations of these parameters in our simulation studies. In particular, we use the following setting:

- Number of simulations($N$): number of simulations is set at 1000.
- Sample size($n$): corresponding to small, medium and large sample sizes, we consider $n = 20,\ 30,\ 40$ and 50.
- Number of biological ($b$) and nested technical ($t$) replicates: we set $b = 3$ with $t = 1$ or $t = 2$, $b = 4$ with $t = 2$ or $t = 3$. So, the total number of replicated measurements ($m_1 = m_2 = m$) available for a gene are 3, 6, 8 and 12, respectively.
- Intermolecular and intramolecular correlations $\rho^{ij}_{g_1g_2},\ \rho^{ij}_{g_1},\ \rho^{ij}_{g_2}$: correlation values are set at three different levels low(L) (range 0.2-0.3), medium(M) (range 0.3-0.5) and clean(H) (range 0.5-0.6).
- We write 'LLL', 'LMH' etc. to denote the range of the true correlation values that we use to simulate replicate data sets. The first letter in the triplet represents the range of intermolecular correlation $\rho^{ij}_{g_1g_2}$ and the remaining two letters represent the ranges of intramolecular correlations $\rho^{ij}_{g_1}$ and $\rho^{ij}_{g_2}$, respectively.

### 2.5.2 Performance Evaluation

In Fig. 2.1, we used a typical experimental design of three biological replicates ($b = 3$) with two technical replicates nested within each biological replicate ($t = 2$) and set the sample size at $n = 20,\ 30,\ 40$ and 50. The horizontal axes in Fig. 2.1 represent the true range of correlation parameters ($\rho^{ij}_{g_1g_2}, \rho^{ij}_{g_1}, \rho^{ij}_{g_2}$) that we used to simulate data. For a few combinations, however, we could not simulate data as the corresponding correlation matrices were not positive definite. The vertical axes represent $\Phi$ ($-\log_2 P$ values) calculated for each blind-case and informed-case model, averaged over 1000 simulations. We use adjacent bars

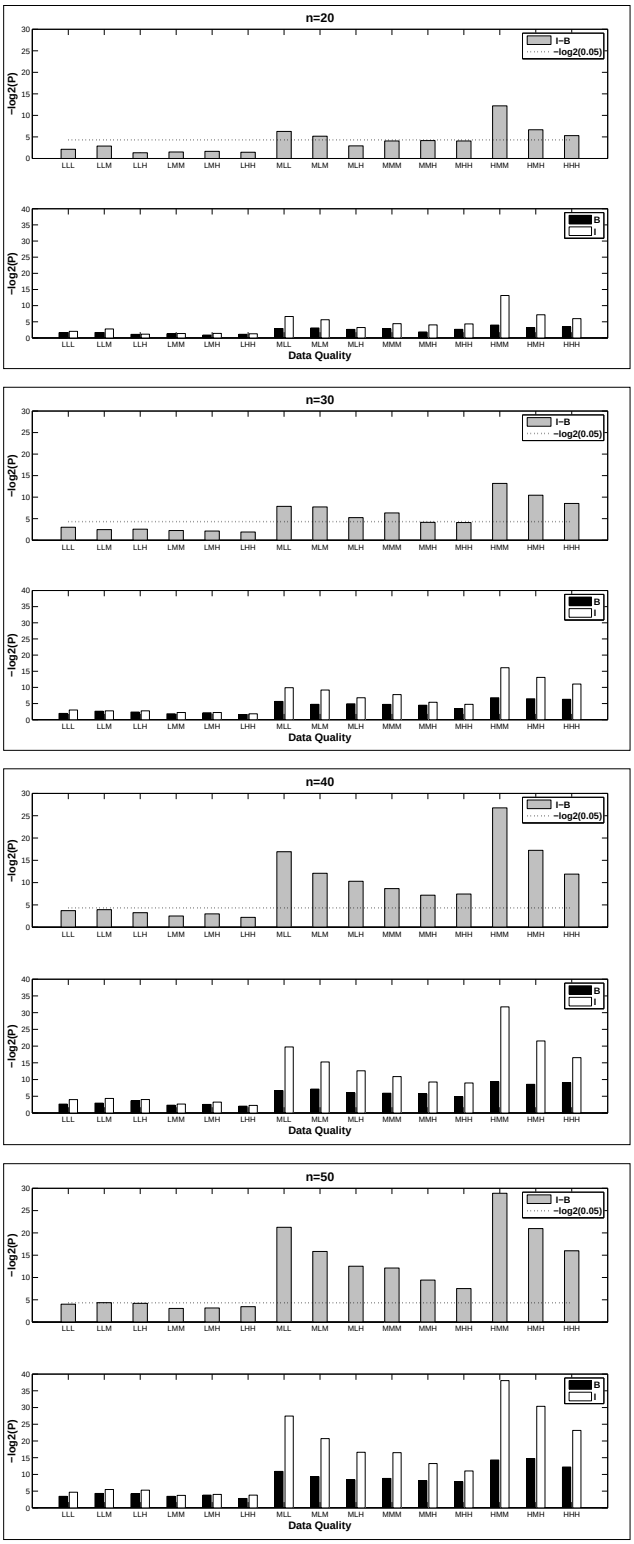


Figure 2.1: Comparison of the blind-case model (B) and the informed-case model (I) using two methods outlined in Section 2.4.3. The simulated data has three biological replicates ($b = 3$) with two technical replicates ($t = 2$) nested within each.

to demonstrate the performances of the informed case (I) and the blind case correlation estimators (B), for a fixed combination of $(\rho^{ij}_{g_1g_2}, \rho^{ij}_{g_1}, \rho^{ij}_{g_2})$.

Clearly, in the lower panels of each of the four blocks in Fig. 2.1, higher $\Phi$ values are observed for the informed-case model. This indicates a better performance of the informed-case model compared with blind-case model. We further tested the statistical significance of the model performances using Method 2 described in Section 2.4.3. In the upper panels of each of the four blocks in Fig. 2.1, we have plotted the $\Phi$ values calculated by comparing $\Psi^I - \Psi^B$ to a chi-square distribution with $J_{m1}J_{m2} - 1$ degrees of freedom. This comparison also demonstrated an overall better performance of the informed-case correlation estimator in the case of replicated measurements with known experimental design.

In Fig. 2.2, we tested the model performances by fixing the sample size at $n = 20$ and setting the numbers of biological and technical replicates at $b = 3, t = 1$, $b = 3, t = 2$, $b = 4, t = 2$ and $b = 4, t = 3$. Both the methods outlined in Section 2.4.3 demonstrated an overall better performance of the informed-case model in comparison to the blind-case model. Moreover, we observed that the contrast between the two models were more pronounced for the small number of replicates: e.g. $b = 3,\ t = 1$, which decreased as the values of $b$ and $t$ increased. Fig. 2.1 and Fig. 2.2 show that the informed-case model outperforms blind-case model the most when the true intermolecular correlation is medium to high, regardless of the data quality. This feature makes the informed-case model particularly useful for predicting functional relationships, which is more meaningful when the biomolecules have medium to high intermolecular correlation.

## 2.6 Discussion

In this chapter, we presented a generalized multivariate model to summarize correlation from replicated and complete molecular profiling data with a prior known replication mechanisms. Since replicated measurements generated from high throughput platforms often have diverse magnitudes, it is necessary to exploit each replicate individually and utilize prior knowledge

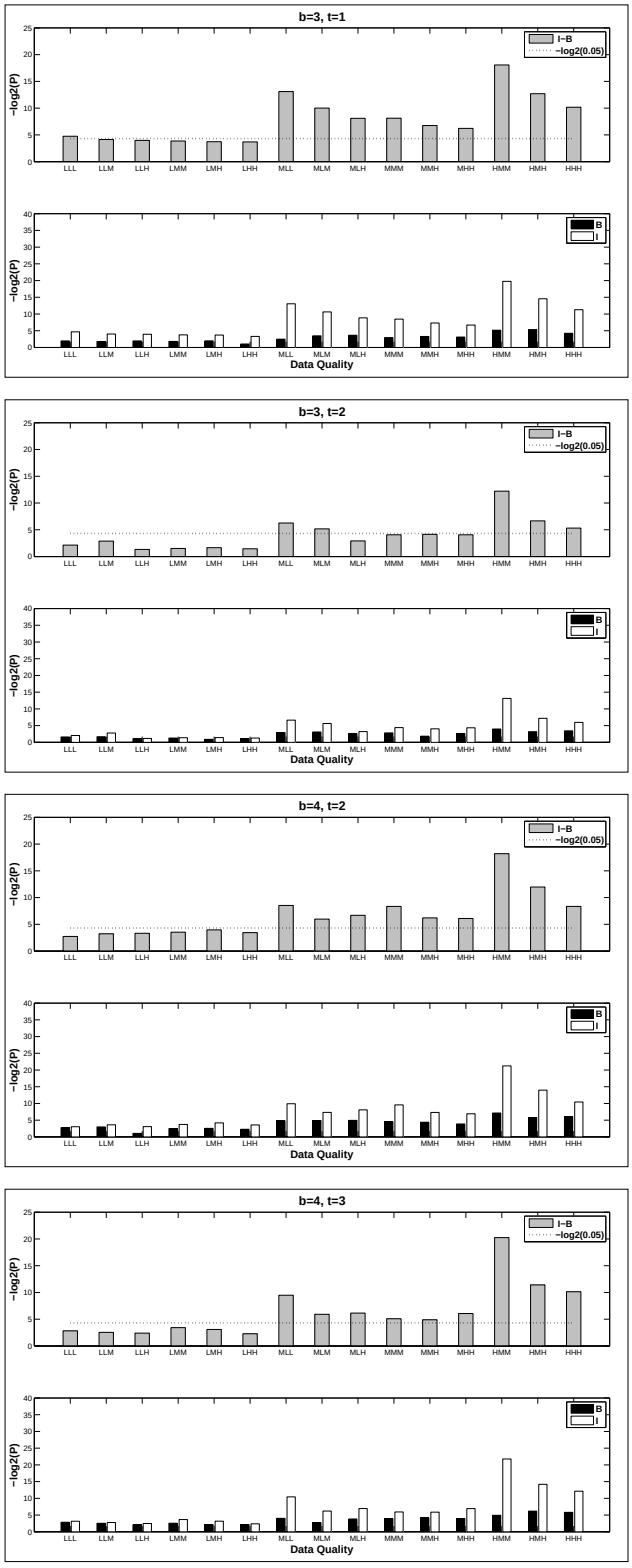


Figure 2.2: Comparison of the blind-case model (B) and the informed-case model (I) with increasing number of biological and technical replicates. Sample size is fixed at $n = 20$.

of replication mechanisms. Traditional bivariate correlation estimators, typically, employ a method of averaging over replicated measurements, which may lead to a significant amount of information loss. In the proposed approach, we treated each replicate as a random variable by assuming that data were random samples from a multivariate normal distribution. The underlying correlation structure was designed to explicitly accommodate prior knowledge of replication mechanisms. We evaluated the performance of our approach by generating replicated data sets with different data quality and between-molecular correlations. The proposed correlation estimator benefits from an easily manageable computational complexity due to the closed-form representations of the MLEs of model parameters. In our analyses, it took only a few seconds to compute the summarization statistics from 1000 different runs of informed-case estimator on a standard desktop computer.

# Chapter 3

# Learning Correlation Structures from Replicated and Complete Molecular Profiling Data II

## 3.1 Introduction

The blind-case and informed-case models presented in the previous chapter estimate the correlation structure from replicated molecular profiling data by imposing constraints on the model parameters. In the case of two genes, for instance, the blind-case correlation structure is defined in terms of three parameters, one within-replicate correlation parameter for each gene and one between-replicate correlation for the gene pair. Informed-case model, on the other hand, introduces more parameters in the correlation structure by assigning different parameters for different biological replicates of the same gene. Such constrained correlation structures, although useful, may suffer from the following limitations: (1) Constrained representations may oversimplify the true correlation structure of a gene set with many pairwise or between-biological replicate correlation structures. (2) Both blind or informed correlation structures may be overly constrained for some data. It is not feasible to design an informed correlation model that will fit for any replicated molecular profiling data. It has been shown in recent researches that the correlation structure underlying molecular profiling data, without simplification, can be estimated reasonably well using shrinkage approaches [125,126,159]. Therefore, it is desirable to develop a flexible approach to adaptively determine the correlation structure of a gene set with replicated and complete measurements.

In this chapter[1], we focus on this problem.

Finite mixture models [36, 90, 91] enable us to estimate such an optimal correlation structure from a mixture of finite number of pre-specified correlation models. The approach assumes that data are independently and identically distributed samples from a mixture of a finite number of distributions with different parameters. Mixture models are often a preferred optimizing tool for solving a wide range of problems when the exact model of the data is hard to discern. Due to their sound mathematical base and interpretability of results, mixture model based approaches have been applied to solve problems in many scientific domains, e.g. [93,94]. we design a two-component finite mixture model by shrinking the correlation structure of a gene set between a model with a constrained set of parameters and the one without any constraints. The proposed mixture model naturally generalizes the constrained model, given by either blind-case or informed-case estimators, presented in Chapter 2. Throughout this chapter, we assume that the constrained component is given by the blind-case estimator.

## 3.2 Notations

Let $G_1, \ldots, G_k$ denote arbitrary biomolecules with $g_{ij}^l$ as their respective abundance levels in the $i^{th}$ replicate (row) and $j^{th}$ sample (column), for $l = 1, \ldots, k$, where the abundance levels are measured over $n$ independent samples. We assume that $m_l$ replicated measurements are available for $G_l$ in each sample, $l = 1, \ldots, k$. The $j^{th}$ column of the given replicated data set is written as $Z_j = (Z_{j1}, \ldots, Z_{jk})^T = (g_{j1}^1, \ldots, g_{jm_1}^1, \ldots, g_{j1}^k, \ldots, g_{jm_k}^k)^T$, $j = 1, 2, \ldots, n$. Data are assumed to be independently and identically distributed samples from a mixture of multivariate normal distributions with $\sum_{l=1}^{k} m_l$ random variables.

[1]Published work [1]. Reused with permission. Copyright, IEEE. All rights reserved.

## 3.3 Finite Mixture Model Approach

### 3.3.1 The Model

In the mixture model approach [36, 90], the density of a sample $Z_j$ is modeled as mixture of a finite number of component densities. In this chapter, we consider the case of two component densities $f_1(Z_j)$ and $f_2(Z_j)$ with mixture proportions $\pi_1$ and $\pi_2$, respectively. For $j = 1, \ldots, n$, this is expressed as

$$f(Z_j, \Theta) = \pi_1 \ f_1(Z_j) + \pi_2 \ f_2(Z_j), \tag{3.1}$$

where $\Theta$ denotes the set of all parameters in the mixture model and $\pi_1 + \pi_2 = 1$.

We consider a mixture of two multivariate normal distributions and denote the parameters for the $i^{th}$ component by $\theta_i = \{\mu_i, \ \Sigma_i\}, \ i = 1, 2$. Thus, we have

$$f_i(Z_j; \theta_i) = \frac{1}{(2\pi)^{\frac{1}{2}(\sum_{l=1}^{k} m_l)n} |\Sigma_i|^{\frac{n}{2}}} \ e^{\{-\frac{1}{2}(Z_j - \mu_i)^T \Sigma_i^{-1} (Z_j - \mu_i)\}} \tag{3.2}$$

for $i = 1, \ 2, \ j = 1, \ldots, n$. We further assume that the first component in the mixture is given by the blind-case (Section 2.3.1) model presented in Chapter 2. This means $(\mu_1, \Sigma_1)$ is $(\mu^B, \Sigma^B)$. The parameters for the second component in the mixture model are free from any constraints.

In the mixture model approach, the posterior probability $(\tau_i)$ that $Z_j$ is sampled from the $i^{th}$ component of the mixture is estimated using an EM algorithm [29, 90]. Each $Z_j$ is considered a sample from a model for which it has higher posterior probability of belonging. For example, the samples $Z_j$'s satisfying $\hat{\tau}_1(Z_j) >= \hat{\tau}_2(Z_j)$ follow the parameter structure of the first component density in the mixture. Here, $\hat{\tau}_i$ denotes the value of $\tau_i, \ i = 1, 2$, estimated using EM algorithm. In the above case, the observations $Z_j$'s with $\hat{z}_{1j} = 1$ are sampled from the first component, where $\hat{z}_j$ is an estimate of the component indicator vector

$z_j,\ j = 1, \ldots, n$. It is defined as

$$\hat{z}_{ij} = (\hat{z}_j)_i = \begin{cases} 1 & \text{if } \hat{\tau}_i(Z_j) >= \hat{\tau}_h(Z_j),\ h \in \{1, 2\},\ h \neq i, \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

### 3.3.2 Unconstrained EM Algorithm

We first discuss the unconstrained EM algorithm which has been used in the literature to perform model based clustering [90]. EM is an iterative procedure which involves two steps: the E step and the M step. Under the EM framework, the observed data ($Z_j$'s in this case) are assumed to be incomplete. The observed $Z_j$'s together with unobserved component-indicator vectors $z_j$'s, represent complete data, $j = 1, 2, \ldots, n$. The E and M steps are described below:

E Step: At the $(k+1)^{th}$ iteration, the E-step computes the conditional expectation of the complete data log likelihood. Complete data log likelihood is given by

$$\log L_c(\Psi) = \sum_{i=1}^{2} \sum_{j=1}^{n} (z_{ij} \log \pi_i + z_{ij} \log f_i(Z_j; \theta_i)). \tag{3.4}$$

The conditional expectation is expressed as

$$Q(\Psi; \Psi^{(k)}) = \sum_{i=1}^{2} \sum_{j=1}^{n} \tau_i(Z_j; \Psi^{(k)})[\log \pi_i + \log f_i(Z_j; \theta_i)]. \tag{3.5}$$

In Eq. 3.5, the symbol $\tau_i(Z_j; \Psi^{(k)})$ represents the posterior probability that $Z_j$ belongs to either first or the second component in the mixture. It is computed using

$$\tau_i(Z_j; \Psi^{(k)}) = \frac{\pi_i^{(k)} f_i(Z_j; \theta_i^{(k)})}{\sum_{h=1}^{2} \pi_h^{(k)} f_h(Z_j; \theta_h^{(k)})}, \tag{3.6}$$

for $i = 1,\ 2$.

M Step: At the $(k+1)^{th}$ iteration, the M-step updates the parameter estimates by maximizing $Q(\Psi; \Psi^{(k)})$ with respect to $\Psi$. The updating rules are given below:

$$\pi_i^{k+1} = \frac{1}{n} \sum_{j=1}^{n} \tau_i(Zj; \Psi^{(k)}) \tag{3.7}$$

$$\mu_i^{k+1} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} Z_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{3.8}$$

$$\Sigma_i^{k+1} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (Z_j - \mu_i^{(k+1)})(Z_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \tag{3.9}$$

for $i = 1,\ 2$, where

$$\tau_{ij}^{(k)} = \tau_i(Z_j; \Psi^{(k)}) \tag{3.10}$$

for each $i = 1,\ 2$ and $j = 1, \cdots, n$.

EM algorithm iteratively proceeds between the E-step and the M-step until convergence. For a more detailed discussion on this topic, we refer to [90].

### 3.3.3 Constrained EM Algorithm

The unconstrained EM algorithm does not guarantee the convergence of the sequence of parameters $\{\Psi^k\}$ towards the MLE $\hat{\Psi}$ for the following reasons

- The generated sequence of log-likelihood values may not be bounded.
- The log-likelihood function may converge to a point of local maximum, which makes the estimation of parameters dependent on the initial guess.

Problems in using unconstrained EM algorithm and their remedies have been reported and investigated in numerous researches [50, 61, 62, 91]. It has been observed that by imposing bounds on the eigenvalues of the component correlation matrices $\Sigma_i, i = 1, 2$, various problems associated with the convergence of EM algorithm can be significantly reduced. In this chapter, we use the constrained EM algorithm from [61] which addresses the above conver-

gence problems by making the eigenvalues of the component correlation matrices lie in a certain interval. The numerical studies presented in [61] demonstrate the convergence of the log-likelihood function to the right maximum in majority of the cases by using constrained EM algorithm. Moreover, the number of successful cases of convergence is higher than that from unconstrained EM algorithm.

The constrained EM algorithm [61] reformulates the constraints considered in [50]. Let $a$ and $b$ be two strictly positive constants satisfying $a/b \geq c$, with $c \in (0\ 1]$. If the eigenvalues of two component correlation matrices satisfy $a \leq \lambda_j(\Sigma_i) \leq b$ for $i = 1, 2, j = 1, 2, \ldots, \sum_{l=1}^{k} m_l$, then the condition $\lambda_{\min}(\Sigma_1 \Sigma_2^{-1}) \geq c$ imposed in [50] is also satisfied and leads to a constrained (global) maximization of the likelihood. Here $\lambda_{\min}$ stands for the smallest eigenvalue. As every symmetric matrix $A$ admits a spectral decomposition $A = SDS^T$, where $D$ is the diagonal matrix formed by the eigenvalues of $A$ and $S$ is an orthogonal matrix whose columns are standardized eigenvectors. In each iteration of the constrained EM algorithm, the eigenvalues of the updated correlation matrices are bound to lie in the interval $[a\ b]$. If an eigenvalue is smaller than $a$, it is replaced by $a$ and if it is is greater than $b$, it is replaced $b$, without changing the matrix of eigenvectors obtained from spectral decomposition.

### 3.3.4 Correlation-Based Clustering

A model estimating the correlation structure can be tested for its performance in revealing feature associations through cluster analysis. The estimated correlation structure can undergo the process of hierarchical clustering equipped with correlation distance metric. In the case when the class labels of data are *a priori* available, the clustering results can be validated by computing Minkowski Scores [65].

Let $C^S$ denote a matrix with $C^S_{ij} = 1$ if the $i^{th}$ and $j^{th}$ feature vectors belong to the same cluster in the solution $S$ obtained by hierarchial clustering and 0 otherwise, and $T$ be the corresponding matrix for the true solution. Then, the Minkowski score corresponding to

the result $S$ is defined as

$$\text{Minkowski score} = \frac{\|C^S - T\|}{\|T\|} \tag{3.11}$$

In model comparisons, a lower Minkowski score implies a better clustering result.

## 3.4 Simulations

### 3.4.1 Simulation Settings

We compare the two-component mixture model and a constrained model comprising of a single component in estimating the true correlation structure. We assume that the first component in the mixture model and constrained model used for comparison are given by the blind-case model (Eq. 2.2). We use mean squared error (MSE)

$$\text{MSE} = \frac{1}{N}\sum_{l=1}^{N}(\sum_{i}\sum_{j}(\Sigma_{ij} - \hat{\Sigma}_{ij}^{(l)})^2)$$

as our criteria to evaluate the model performances. Here $\Sigma$ and $\hat{\Sigma}$ denote the true and estimated correlation structures, respectively, where $\Sigma = \pi_1\Sigma_1 + \pi_2\Sigma_2$ with $\pi_i$ and $\Sigma_i,\ i = 1, 2$ as true mixture proportions and the component correlation matrices, respectively. $N$ is the total number of simulations and $\hat{\Sigma}^{(l)}$ is the $l^{th}$ estimate of $\Sigma$. After the convergence of EM algorithm, $\hat{\mu}_1$ and $\hat{\Sigma}_1$ are assigned their constrained structure, as presented in Eq. 2.1 and Eq. 2.2, by averaging their component blocks. For instance, the first $m_1$ entries in $\hat{\mu}_1$ are replaced by their average. However, $\hat{\mu}_2$ and $\hat{\Sigma}_2$ remain the same as obtained using EM. The convex combination of $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$, with estimated mixture proportions as coefficients, determines $\hat{\Sigma}$. This means $\hat{\Sigma} = \hat{\pi}_1\hat{\Sigma}_1 + (1 - \hat{\pi}_2)\hat{\Sigma}_2$.

We show that mixture model outperforms the constrained model by considering realistic combinations of various parameters. In particular, we set

- Number of simulations ($N$): we fix the number of simulation at $N = 1000$.

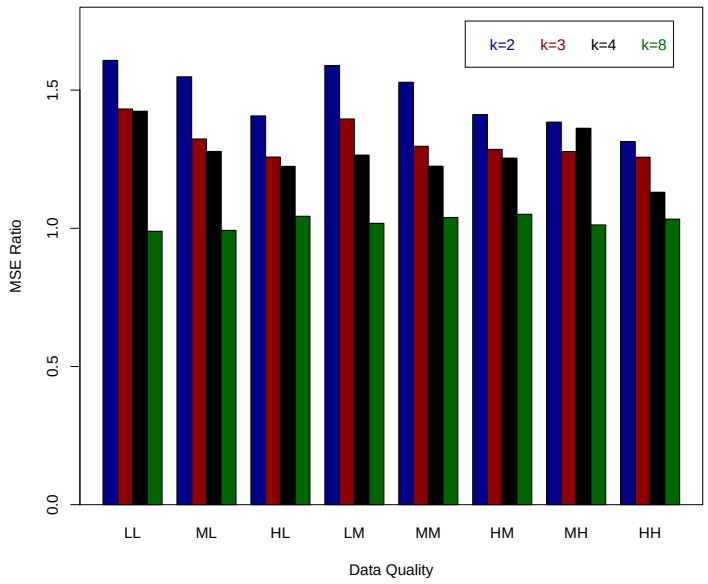


Figure 3.1: Comparison of the mixture model and the blind-case model in terms of MSE ratio, where MSE ratio = MSE from the blind-case model/MSE from the mixture model.

- Number of samples ($n$): we set the number of samples at $n = 20$.
- Number of genes ($k$): it is the number of genes which form replicated data. It is set at four different values $k = 2, 3, 4$ and 8.
- Number of replicates ($m_l$): as most genome-wise data have a few replicates due to high experimental costs, we set $m_l = m = 4$, for $l = 1, \ldots, k$.
- Correlation parameters ($\rho_{ij}$) (see Section 2.3.1): in the case of a constrained model (blind-case model in this case), within-molecular correlation ($\rho_{ii}$) and between-molecular correlation ($\rho_{ij}$) values are set at three different levels, low(L)(0.2-0.3), medium(M)(0.3-0.5) and clean(H)(0.5-0.6).
- We consider different combinations of the correlation parameters to simulate data and express it by writing the pairs (L,M), (M,H) etc., where the first entry corresponds to the range of between-molecular correlation and the second entry denotes within-molecular correlation range.

### 3.4.2 Performance Evaluation

In Fig. 3.1, we compare the model performances in terms of MSE ratio, which is the ratio of MSE from the blind-case model over the one from mixture model. In the figure, the horizontal axis represents the range of true correlation parameters and the vertical axis represents MSE ratios. Ratios greater than 1 indicate that mixture model outperforms the blind-case model. We compared the two estimators by varying the range of correlation parameters from (L,L) to (H,H) and setting the number of genes at $k = 2, 3, 4$ and 8. In Fig. 3.1, almost all examined MSE ratios are greater than 1 which indicates the superior performance of the mixture model approach compared with blind-case model in estimating the population correlation structure. As the blind-case model is an increasing function of the number of feature vectors in replicated data [158], MSE ratio decreases with increase in the number of genes. For a few combinations of parameters, simulations could not be preformed as the corresponding correlation matrices were not positive definite.

## 3.5 Real-world Data Analysis

### 3.5.1 Data

The model performances were evaluated using two publically available replicated data sets

- spike-in data from Affymetrix (`http://www.affymetrix.com`) and
- yeast galactose data from [152] (`http://expression.washington.edu/publications/kayee/yeunggb2003/`).

Affymetrix has made spike-in data sets available as benchmark to compare different probe set expression summarization methods, such as RMA [64] and GCRMA [150]. We use spike-in data as benchmark to compare the estimated correlation structure with the nominal correlation structure. Spike-in data consists of the expression levels of 16 genes, each with 16 replicated measurements. For spike-in data, nominal pairwise correlation values can be

obtained from known probe-level intensities. As a special case, we compare the pairwise correlations estimated from the mixture model and blind-case model with the nominal correlation. The yeast galactose data contains the expression levels of 205 genes, each with 4 replicated measurements. Yeast data set was used to test the clustering performance of models, which could be assessed from the class labels of genes available *a priori*. Indeed, the 205 genes were previously classified into four functional groups [152]. The correlation structures estimated using the blind-case model and mixture model were used to perform hierarchial clustering by employing correlation as a distance metric. Clustering performance of each model was assessed in terms of Minkowski score.

### 3.5.2 Estimation of Correlation Structure

In the following steps, we summarize the procedure to choose the initial values in EM algorithm

- For the unconstrained component, the initial mean vector is chosen as the sample mean. For the constrained component, the sample mean is given a constrained structure by averaging its component blocks.

- For the unconstrained component, the initial correlation matrix is obtained by computing all pairwise Pearson's correlations. For the constrained component, we use a constrained structure obtained by averaging the component blocks of unconstrained correlation structure.

- Values of $a$ and $b$ are taken to be the minimum and maximum eigenvalues of the initial correlation matrices, respectively.

In Fig. 3.2, we compare the squared error values in estimating pairwise correlations from the mixture model and blind-case model using spike-in data set. Here, the $x$-axis represents different probe pairs and the $y$-axis denotes the squared error values from the two models. It was observed that in almost 82% cases, mixture model showed a lower squared

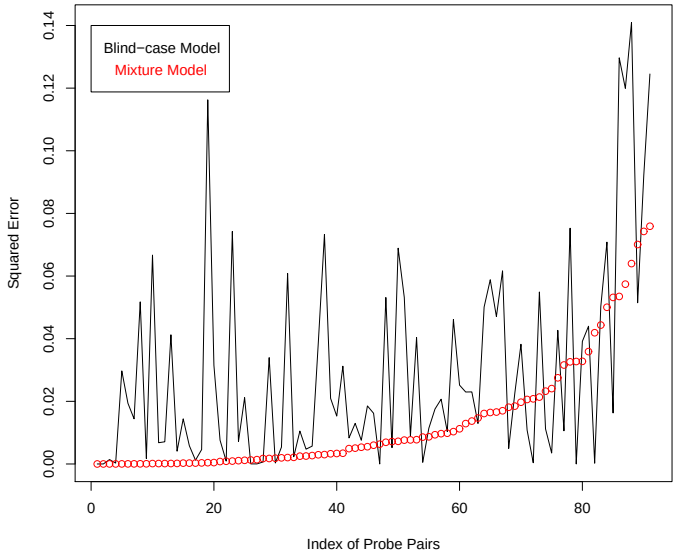


Figure 3.2: Comparison of the squared error values in estimating all pairwise correlations using the mixture model and the blind-case model, for spike-in data

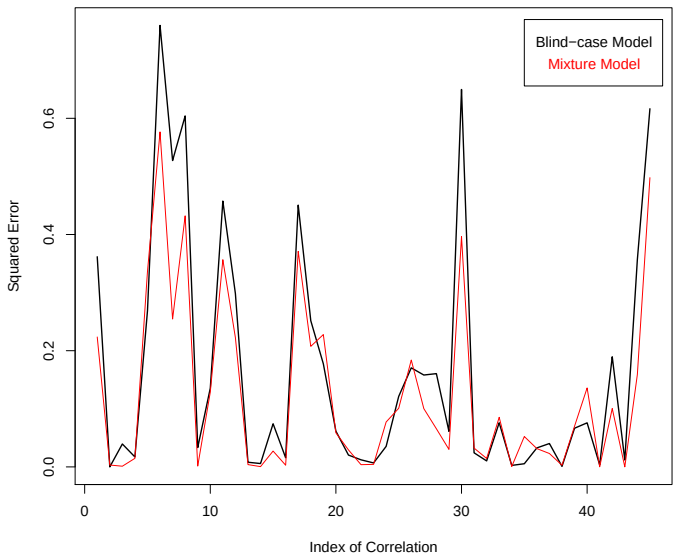


Figure 3.3: Comparison of the correlation structures estimated using the mixture model and the blind-case model with the nominal correlation structure, for selected probe sets in spike-in data

error in estimating pairwise correlation than the blind-case model. Further, the performance of mixture model could be enhanced up to 89% by discarding a couple of probes from the analysis.

As dimension is often an issue in estimating the correlation structure, we considered 10 selected probe sets from spike-in data and compared the two correlation structures estimated using mixture model and blind-case model with the nominal correlation structure. In Fig. 3.3, we have plotted the squared error values in estimating the correlation structure. It is clear from Fig. 3.3 that in almost all the cases, the squared error values in the case of mixture model were lower than the ones from the blind-case model.

### 3.5.3 Cluster Analysis

Fig. 3.4 demonstrates the model performances in terms of Minkowski score for 150 different subsets of the yeast data, each of which correspond to 60 randomly selected probe sets. In Fig. 3.4, the $x$-axis denotes the index of the selected subset and the $y$-axis plots the Minkowski scores from the blind-case model and mixture model. A better performance of the mixture model was supported by lower Minkowski scores compared with the constrained model, in almost 73% cases.

Thus, we claim that our two-component mixture model based approach leads to (1) a better estimation of the true correlation structure possessing lower squared error values and (2) better clustering results with lower Minkowski score, in comparison to the one component blind-case model.

## 3.6 Discussion

We adopted a two-component mixture model approach to estimate the correlation structure of a gene set from replicated and complete molecular profiling data. We assumed that data are independently and identically distributed samples from a mixture of two multivariate normal distributions, one with a constrained and the other with an unconstrained param-

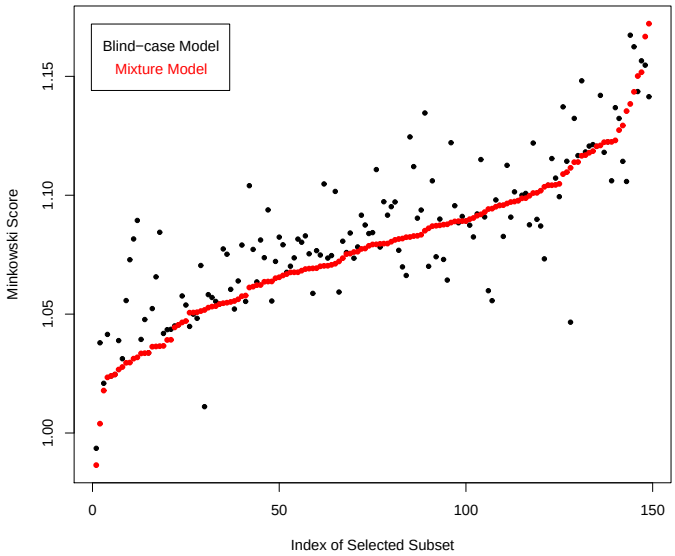


Figure 3.4: Performance of the blind-case model and the mixture model in clustering yeast data. Each index corresponds to a data set with 60 randomly selected probe sets.

eter structure. In our analyses, the constrained component in the mixture was given by the blind-case model. Simulation and real-world data analysis showed that mixture model based estimator possessed an overall lower (mean) squared error in estimating correlation structure and lower Minkowski score in clustering, than the constrained model given by the first component in the mixture. This clearly indicates that the two-component mixture model proposed in this chapter is better in estimating the overall correlation structure from molecular profiling data. Apart from clustering, our approach can be used in various analysis techniques for pattern discovery which require a reliable estimation of correlation including gene association networks, classification methods, e.g. linear and quadratic discriminant analysis.

# Chapter 4

# Learning Correlation Structures from Replicated and Incomplete Molecular Profiling Data

## 4.1 Introduction

In Chapters 2-3, our focus was on a correlation based pattern discovery from replicated and complete molecular profiling data. However, the expression profiles generated from high throughput experiments are not only replicated, they often contain a small to large percentage of missing values [81]. For replicated and incomplete molecular profiles of diverse magnitude, missing value imputation by sample mean or median could be strongly biased by high ratios of missing values and/or low data quality. The commonly used k-nearest neighbor data imputation algorithm [146] often fails when a missing column has no neighbor with complete measurements. There is an urgent need to develop new approaches to fully exploit the replicated and incomplete genome-wide measurements. In this chapter[1], we present a generalized multivariate approach, developed under the Expectation-Maximization framework, to estimate the underlying correlation structure from replicated and incomplete molecular profiling data with either blind-case or an informed-case replication mechanisms.

## 4.2 Notations

Throughout this chapter, $G_1, \ldots, G_k$ denote arbitrary biomolecules with $g_{ij}^l$ as their respective abundance levels in the $i^{th}$ replicate and $j^{th}$ sample, for $l = 1, \ldots, k$, where the

[1]Published work [161]. Reused with permission. Copyright, IEEE. All rights reserved.

abundance levels are measured over $n$ independent samples. Further, we assume that $m_l$ replicated measurements are available for $G_l$ in each sample, $l = 1, \ldots, k$. The $j^{th}$ column of the given data $Z_j = (Z_{j1}, \ldots, Z_{jk})^T = (g^1_{j1}, \ldots, g^1_{jm_1}, \ldots, g^k_{j1}, \ldots, g^k_{jm_k})^T$, $j = 1, 2, \ldots, n$, and is assumed to be an independently and identically distributed sample from a multivariate normal distribution, given by either blind-case $N(\mu^B, \Sigma^B)$ or informed-case model $N(\mu^I, \Sigma^I)$, with $\sum_{l=1}^{k} m_l$ random variables. Other notations related to the two models are borrowed from Chapter 2.

## 4.3 EM Algorithm

In this section, we present a novel EM algorithm to estimate the underlying correlation structure from replicated and incomplete molecular profiling data. Without loss of generality, we assume that data are sampled from a multivariate normal distribution with informed-case correlation structure. The algorithm for the blind-case follows as a particular case. For simplicity, we present our algorithm in the case of two genes $G_1$ and $G_2$. Recalling from Chapter 2, the mean vector $\mu^I$ for the informed-case model is defined in terms of the scalars $\mu^i_{g_1}$ and $\mu^i_{g_2}$, and $\Sigma^I$ is defined in terms of the correlation parameters $\rho^{ij}_{g_1 g_2}$, $\rho^{ij}_{g_1}$ and $\rho^{ij}_{g_2}$, where $i$ and $j$ represent indices for the biological replicates of the genes $G_1$ and $G_2$, respectively.

It is well-known that the sufficient statistics for the multivariate normal distribution are,

$$T_1 = \sum_{j=1}^{n} Z_j \tag{4.1}$$

and

$$T_2 = \sum_{j=1}^{n} Z_j Z_j^T = Z Z^T \tag{4.2}$$

where $Z = (Z_1, Z_2, \ldots, Z_n)$ is the matrix comprising of all samples. We utilize the above statistics in the E and M steps of the proposed EM algorithm.

### 4.3.1 The E Step

Without loss of generality, we can assume that column $Z_i,\ 1 \leq i \leq r$ has missing values, for a fixed number $r$, $1 \leq r \leq n$. We denote a missing entry by prefixing it with the letter $M$. Let us consider a particular case where a column $Z_j$ with missing values is of the form

$$\begin{aligned} Z_j &= (Mg^1_{1j}, \ldots, Mg^1_{I^1_{m_1}j}, g^1_{(I^1_{m_1}+1)j}, \ldots, g^1_{m_1 j}, Mg^2_{1j}, \ldots, Mg^2_{I^1_{m_2}j}, g^2_{(I^1_{m_2}+1)j}, \ldots, g^2_{m_2 j})^T \\ &= (MG^j_1, cG^j_1, MG^j_2, cG^j_2)^T \end{aligned} \tag{4.3}$$

for $1 \leq j \leq r$. In this illustration we have assumed that replicated measurements corresponding to the first biological replicate in the two genes $G_1$ and $G_2$ are missing, which can be generalized to the case of columns with missing values in any location. Here $MG^j_1$ represents the segment with missing entries corresponding to the first biological replicate, in the $j^{th}$ sample of $G_1$. Further, $cG^j_1$ represents the 'complementary' segment in the $j^{th}$ sample of $G_1$ with no missing entries. A similar explanation holds good for $MG^j_2$ and $cG^j_2$.

For random initial guesses $\mu^{I^{(0)}}$ and $\Sigma^{I^{(0)}}$, in the $(t+1)^{th}$ iteration, the E-step computes the expected value of $T_1$ and $T_2$ in the presence of observed data $Z_{obs}$ and current estimate of parameters $\mu^{I^{(t)}}$ and $\Sigma^{I^{(t)}}$ as follows

$$\begin{aligned} E(T_1 | Z_{obs}, \mu^{I^{(t)}}) &= E(\sum_{j=1}^{r} Z^{(t)}_j) + \sum_{j=r+1}^{n} Z^{(t)}_j \\ &= A^{(t)} + \sum_{j=r+1}^{n} Z^{(t)}_j \end{aligned} \tag{4.4}$$

where

$$A^{(t)} = \begin{pmatrix} r\mu_{g_1}^{1\ (t)} e_{(I^1_{m_1})} \\ \sum_{j=1}^{r} cG_1^j \\ r\mu_{g_2}^{1\ (t)} e_{(I^1_{m_2})} \\ \sum_{j=1}^{r} cG_2^j \end{pmatrix}$$

where $e_{(p)}$ denotes $p$-length vector $(1, 1, \ldots, 1)^T$. Computationally, $A^{(t)}$ is obtained as below:

- Replace all missing entries corresponding to $j_{m_1}^{th}$ biological replicate in $G_1$ by $\mu_{g_1}^{j_{m_1}(t)}$, for $1 \leq j_{m_1} \leq J_{m_1}$.
- Replace all missing entries corresponding to $j_{m_2}^{th}$ biological replicate in $G_2$ by $\mu_{g_2}^{j_{m_2}(t)}$, for $1 \leq j_{m_2} \leq J_{m_2}$.
- Sum up the first $r$ columns to form $A^{(t)}$.

Further, the columns imputed above are used to form the matrices $M_j'^{(t)},\ j = 1, 2, \ldots, r$, as follows

- Form the matrices $M_j^{(t)} = Z_j^{(t)} Z_j^{(t)T}$, $j = 1, 2, \ldots, r$.
- In $M_j^{(t)}$, replace the entries $\mu_{g_1}^{j_{m_1}(t)} \times \mu_{g_1}^{k_{m_1}(t)}$ by $\mu_{g_1}^{j_{m_1}(t)} \times \mu_{g_1}^{k_{m_1}(t)} + \rho_{g_1}^{j_{m_1}k_{m_1}(t)}$ for gene $G_1$, where $1 \leq j_{m_1}, k_{m_1} \leq J_{m_1}$. An analogous modification is done for gene $G_2$.
- Replace the entries $\mu_{g_2}^{j_{m_1}(t)} \times \mu_{g_2}^{j_{m_2}(t)}$ by $\mu_{g_2}^{j_{m_1}(t)} \times \mu_{g_2}^{j_{m_2}(t)} + \rho_{g_1 g_2}^{j_{m_1} j_{m_2}(t)}$, for $1 \leq j_{m_1} \leq J_{m_1}$ and $1 \leq j_{m_2} \leq J_{m_2}$.
- Matrices obtained from $M_j^{(t)}$ after making above replacements are $M_j'^{(t)},\ j = 1, 2, \ldots, r$.

Then,

$$\begin{aligned} E(T_2 | Z_{obs}, \mu^{I^{(t)}}, \Sigma^{I^{(t)}}) &= E(\sum_{j=1}^{r} Z_j^{(t)} Z_j^{(t)T} | Z_{obs}, \mu^{I^{(t)}}, \Sigma^{I^{(t)}}) + \sum_{j=r+1}^{n} Z_j^{(t)} Z_j^{(t)T} \\ &= \sum_{j=1}^{r} M_j'^{(t)} + \sum_{j=r+1}^{n} Z_j^{(t)} Z_j^{(t)T}. \end{aligned} \tag{4.5}$$

If we write the matrix $M_j^{\prime(t)}$ as

$$M_j^{\prime(t)} = \begin{pmatrix} M_{j_{g_1}}^{\prime(t)} & M_{j_{g_1g_2}}^{\prime(t)} \\ M_{j_{g_1g_2}}^{\prime(t)T} & M_{j_{g_2}}^{\prime(t)} \end{pmatrix}$$

then in case of Equation 4.3, we have

$$M_{j_{g_1}}^{\prime(t)} = \begin{pmatrix} (1-\rho_{g_1}^{11(t)})E_{I_{m_1}^1} + R_{g_1}^{11(t)} e_{(I_{m_1}^1)} {e_{(I_{m_1}^1)}}^T & {\mu_{g_1}^1}^{(t)} e_{(I_{m_1}^1)}\ {cG_1^j}^T \\ cG_1^j\ {\mu_{g_1}^1}^{(t)} e_{(I_{m_1}^1)}^T & cG_1^j\ {cG_1^j}^T \end{pmatrix}$$

$$M_{j_{g_2}}^{\prime(t)} = \begin{pmatrix} (1-\rho_{g_2}^{11(t)})E_{I_{m_2}^1} + R_{g_2}^{11(t)} e_{(I_{m_2}^1)} {e_{(I_{m_2}^1)}}^T & {\mu_{g_2}^1}^{(t)} e_{(I_{m_2}^1)}\ {cG_2^j}^T \\ cG_2^j\ {\mu_{g_2}^1}^{(t)} e_{(I_{m_2}^1)}^T & cG_2^j\ {cG_2^j}^T \end{pmatrix}$$

and

$$M_{j_{g_1g_2}}^{\prime(t)} = \begin{pmatrix} ({\mu_{g_1}^1}^{(t)} {\mu_{g_2}^1}^{(t)} + {\rho_{g_1g_2}^{11}}^{(t)})\ e_{(I_{m_1}^1)} {e_{(I_{m_2}^1)}}^T & {\mu_{g_1}^1}^{(t)} e_{(I_{m_1}^1)}\ {cG_2^j}^T \\ cG_2^j\ {\mu_{g_1}^1}^{(t)} e_{(I_{m_1}^1)}^T & cG_1^j\ {cG_2^j}^T \end{pmatrix}.$$

where $E_{I_{m_1}^1}$ and $E_{I_{m_2}^1}$ are identity matrices of order $I_{m_1}^1$ and $I_{m_2}^1$, respectively and

$$R_{g_1}^{11(t)} = {\mu_{g_1}^1}^{(t)} {\mu_{g_1}^1}^{(t)} + \rho_{g_1}^{11(t)}$$

$$R_{g_2}^{11(t)} = {\mu_{g_2}^1}^{(t)} {\mu_{g_2}^1}^{(t)} + \rho_{g_2}^{11(t)}.$$

It is important to specify here that, only the blocks $M_{j_{g_1g_2}}^{\prime(t)}$ participate in estimating correlation.

### 4.3.2 The M Step

For complete case, the MLE of $\mu^I$ is $\hat{\mu}^I = \sum_{j=1}^{n} Z_j/n$, and the MLE of $\Sigma$ is $\hat{\Sigma}^I = n^{-1}ZZ^T - \hat{\mu}^I\hat{\mu}^{I^T}$. Since, the E step makes the data complete, we write

$$\mu^{I^{(t+1)}} = E(\sum_{j=1}^{n} Z_j | Z_{obs}, \mu^{I^{(t)}})/n = (A^{(t)} + \sum_{j=r+1}^{n} Z_j)/n \tag{4.6}$$

and,

$$\begin{aligned} \Sigma^{I^{(t+1)}} &= n^{-1}E(T_2|Z_{obs}, \mu^{I^{(t)}}, \Sigma^{I^{(t)}}) - \mu^{I^{(t+1)}}\mu^{I^{(t+1)^T}} \\ &= n^{-1}\{\sum_{j=1}^{r} M_j'^{(t)} + \sum_{j=r+1}^{n} Z_j Z_j^T\} - \mu^{I^{(t+1)}}\mu^{I^{(t+1)^T}}. \end{aligned} \tag{4.7}$$

The next iterates of parameters are obtained by averaging the component blocks $\Sigma_{uv}^{ij},\ u, v \in \{g_1, g_2\}$ (see Eq. 2.11) containing the measurements for the $i^{th}$ and $j^{th}$ biological replicate in $\Sigma^{I^{(t+1)}}$ and component blocks of $\mu^{I^{(t+1)}}$ as follows

$$\begin{aligned} \rho_{g_1g_2}^{ij^{(t+1)}} &= Avg(\Sigma_{g_1g_2}^{ij^{(t+1)}}) \\ \rho_{g_1}^{ij^{(t+1)}} &= Avg(\Sigma_{g_1}^{ij^{(t+1)}}) \\ \rho_{g_2}^{ij^{(t+1)}} &= Avg(\Sigma_{g_2}^{ij^{(t+1)}}) \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} \mu_{g_1}^{i^{(t+1)}} &= Avg(\mu_{\sum_{l=1}^{i} I_{m_1}^{l-1}+1}^{I^{(t+1)}}, \ldots, \mu_{\sum_{l=1}^{i} I_{m_1}^{l}}^{I^{(t+1)}}) \\ \mu_{g_2}^{i^{(t+1)}} &= Avg(\mu_{\sum_{l=1}^{i} I_{m_2}^{l-1}+1}^{I^{(t+1)}}, \ldots, \mu_{\sum_{l=1}^{i} I_{m_2}^{l}}^{I^{(t+1)}}). \end{aligned} \tag{4.9}$$

In the special case of the parsimonious model, we have

$$\begin{aligned}\rho^{(t+1)} &= Avg(\Sigma^{B\ (t+1)}_{g_1g_2})\\ \rho^{(t+1)}_{g_1} &= Avg(\Sigma^{B\ (t+1)}_{g_1})\\ \rho^{(t+1)}_{g_2} &= Avg(\Sigma^{B\ (t+1)}_{g_2})\end{aligned} \tag{4.10}$$

and

$$\begin{aligned}\mu^{(t+1)}_{g_1} &= Avg(\mu_1^{B^{(t+1)}}, \ldots, \mu_{m_1}^{B^{(t+1)}})\\ \mu^{(t+1)}_{g_2} &= Avg(\mu_{m_1+1}^{B^{(t+1)}}, \ldots, \mu_{m_1+m_2}^{B^{(t+1)}})\end{aligned} \tag{4.11}$$

The algorithm iterates between the $E$ and $M$ steps until $\|\mu^{I(t)} - \mu^{I(t+1)}\| + \|\Sigma^{I(t)} - \Sigma^{I(t+1)}\|$ (or $\|\mu^{B(t)} - \mu^{B(t+1)}\| + \|\Sigma^{B(t)} - \Sigma^{B(t+1)}\|$ for blind case estimator) is smaller than a pre-specified constant. This gives the MLE's $\hat{\mu}^I$ and $\hat{\Sigma}^I$ (or $\hat{\mu}^B$ and $\hat{\Sigma}^B$). The informed case estimator (or blind case estimator) quantifies $\hat{\rho}$ by averaging $\hat{\Sigma}^I_{g_1g_2}$ (or $\hat{\Sigma}^B_{g_1g_2}$).

## 4.4 Simulations

### 4.4.1 Simulation Settings

We demonstrate the performance of EM algorithm using both blind-case (Eq. 2.2) and informed-case (Eq. 2.11) models. We use mean squared error (MSE) to judge the performance of various correlation estimation methods. MSE in estimating $\rho$ is defined as

$$MSE_\rho = \sum_{i=1}^{N} (\hat{\rho} - \rho)^2 / N,$$

where $N$ is total number of simulations, $\rho$ represents true intermolecular correlation and $\hat{\rho}$ is the MLE of $\rho$ estimated using any of the following methods.

### Multivariate Methods

We use the following methods, each of explicitly consider each replicate:

- **EM:** Estimates $\hat{\rho}$ using the iterative EM algorithm introduced in Section 4.3 for incomplete replicated data.
- **Mean:** Imputes missing data in a row, by row mean and estimates $\hat{\rho}$ using the blind-case (Eq. 2.2) and informed-case correlation estimators (Eq. 2.11).
- **Med:** Imputes missing values in a row, by row median and estimates $\hat{\rho}$ using the blind-case (Eq. 2.2) and informed-case correlation estimators (Eq. 2.11).
- **KNN:** Imputes missing values using k-nearest neighbor algorithm [146] and estimates $\hat{\rho}$ using the blind-case (Eq. 2.2) and informed-case correlation estimators (Eq. 2.11). (See Appendix A.3).

### Bivariate Methods

We used the methods based on averaging or summarizing over replicated measurements:

- **MeanPear:** Imputes missing values in a row, by row mean and computes Pearson's correlation by averaging over replicates.
- **MeanWtd:** Imputes missing values in a row, by row mean and computes Standard Deviation (SD)-weighted correlation [59] (see Appendix A.4).
- **MedPear:** Imputes missing values in a row, by row median and computes Pearson's correlation by averaging over replicates.
- **MedWtd:** Imputes missing values in a row, by row median and computes SD-weighted correlation.
- **Pear:** Uses complete data set and computes Pearson's correlation by averaging over replicates.

- **Wtd:** Computes SD-weighted correlation from complete data set.

We set other parameters as follows:

- Number of simulations $N = 1000$.
- Sample size as $n = 20$.
- We assume $m_1 = m_2 = m$. Number of replicates $m = 4$ for the blind-case model and $m = 6$ for the informed-case model. For the informed-case model, 6 replicated measurements corresponding to a biomolecule are obtained by generating data sets with 3 biological replicates and 2 technical replicates nested within each.
- Intramolecular correlations $\rho_{g_1}$ (or $\rho_{g_1}^{ij}$) and $\rho_{g_2}$ (or $\rho_{g_2}^{ij}$) are set at three different levels low(L)$(0.1 - 0.3)$, medium (M)$(0.3 - 0.5)$ and clean(H)$(0.5 - 0.6)$.
- Intermolecular correlation values $\rho$ (or $\rho_{g_1g_2}^{ij}$) are set at three levels low(L)$(0.2 - 0.3)$, medium(M)$(0.3 - 0.5)$ and clean$(0.5 - 0.6)$.
- The triplet ground truth LLL, MHL etc. represent the range of true correlations ($\rho$, $\rho_{g_1}$ and $\rho_{g_2}$) or ($\rho_{g_1g_2}^{ij}$, $\rho_{g_1}^{ij}$ and $\rho_{g_2}^{ij}$) used to simulate data.
- Percentage of missing data is set at 6 different levels, 5%, 10%, 15%, 20%, 25% and 30%.
- Initial guesses of parameters $\mu_{g_1}^{i}$ and $\mu_{g_2}^{j}$ are obtained by averaging data corresponding to $i^{th}$ and $j^{th}$ biological replicate in $G_1$ and $G_2$ respectively. For the blind-case, it corresponds to the average of all measurements in $G_1$ and $G_2$ respectively.
- $\rho_{g_1g_2}^{ij}$, $\rho_{g_1}^{ij}$ and $\rho_{g_2}^{ij}$ (or $\rho$, $\rho_{g_1}$ and $\rho_{g_2}$) are assigned arbitrary initial values in the range 0.1-0.3, since real-world data are often noisy.

- A simulation runs until

$$\sum_{i=1}^{2m}(\mu_i^{(t)} - \mu_i^{(t+1)})^2 + \sum_{i=1}^{2m}\sum_{j=1}^{2m}(\Sigma_{ij}^{(t)} - \Sigma_{ij}^{(t+1)})^2 < 10^{-20},$$

  where $\Sigma$ and $\mu$ represent correlation matrix and mean vector for either blind-case or informed-case estimator.

### 4.4.2 Performance Evaluation

We first analyzed the performance of EM algorithm with increasing data quality and percentage of missing values using both blind-case (Eq. 2.2) and informed-case (Eq. 2.11) models. Fig. 4.1 demonstrates these two prospects. In Fig. 4.1, horizontal axis represents the combinations $(\rho, \rho_{g_1}, \rho_{g_2})$ (blind-case) or $(\rho^{ij}_{g_1g_2}, \rho^{ij}_{g_1}, \rho^{ij}_{g_2})$ (informed-case) used to generate data. In Fig. 4.2, horizontal axis represents the percentage of missing values ranging from 5% to 30%. The vertical axis in the two figures plots MSE values. It is clear from Fig. 4.1 and Fig. 4.2 that the performance of EM algorithm is not sensitive to the percentage of missing data when the intermolecular correlation is low. As the intermolecular correlation increase, MSE increases and is more sensitive to the percentage of missing values. However, the EM algorithm is not sensitive to the data quality when the intermolecular correlation is fixed. The insensitivity to data quality when intermolecular correlation is fixed makes the EM algorithm a robust approach to an array of real-world bioinformatics problems.

In Figs. 4.3-4.4 and Figs. 4.5-4.6, we compare the performance of EM with three other multivariate models: Mean, Med and KNN in terms of MSE ratio i.e. the ratio of MSE from Mean, Med or KNN over MSE from EM. A ratio more than 1 indicates better performance of EM method. Figs. 4.3-4.4 correspond to the blind-case model and Figs. 4.5-4.6 correspond to the informed-case model for data sets with 5%, 10%, 15%, 20%, 25% and 30% missing values. For data sets with more than 15% missing values, simulation problems occur with KNN method because the $k^{th}$ nearest neighbors does not often exist,

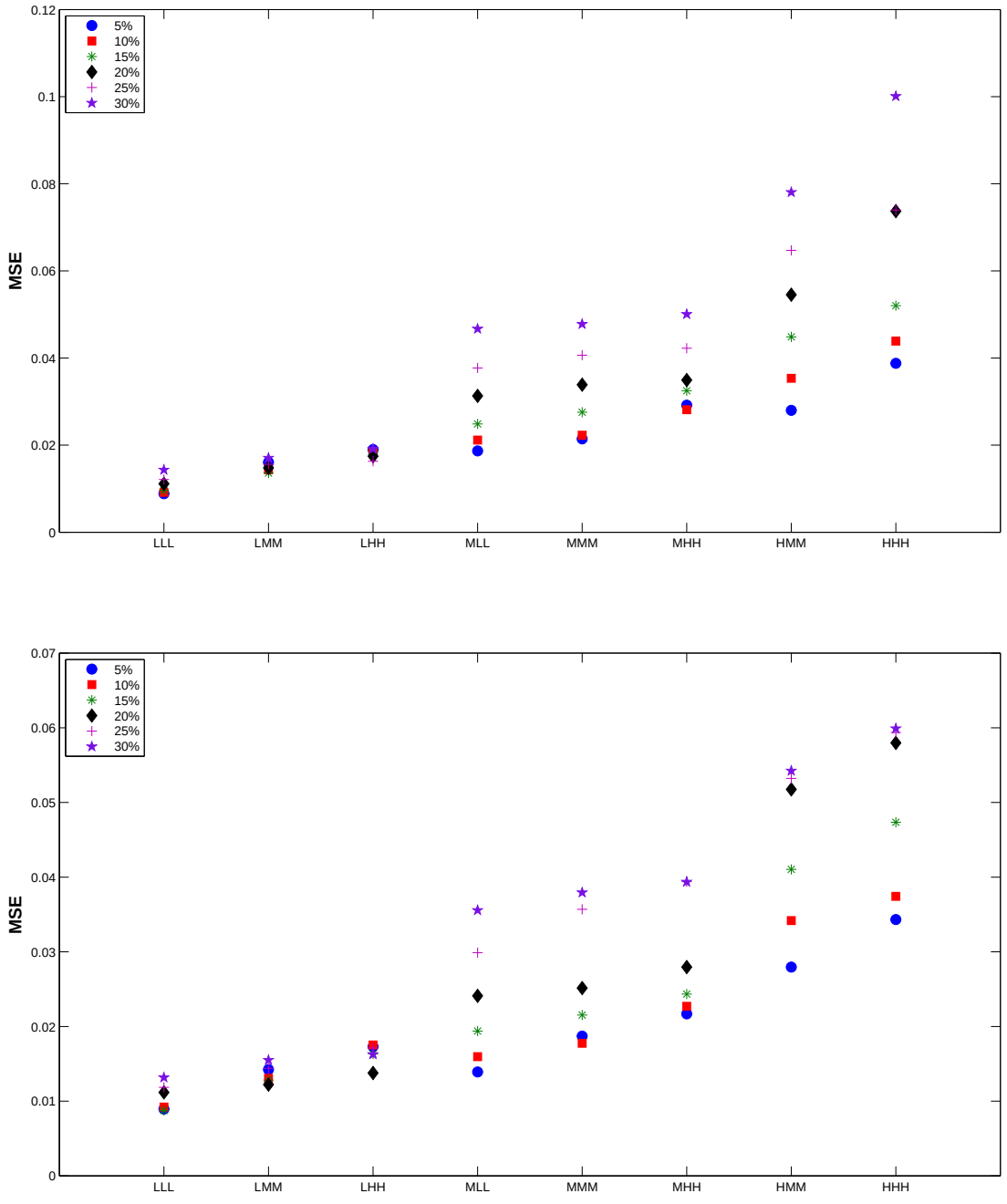


Figure 4.1: Performance of the EM algorithm with increasing data quality. Upper Panel: Blind-case model; Lower Panel: Informed-case model.

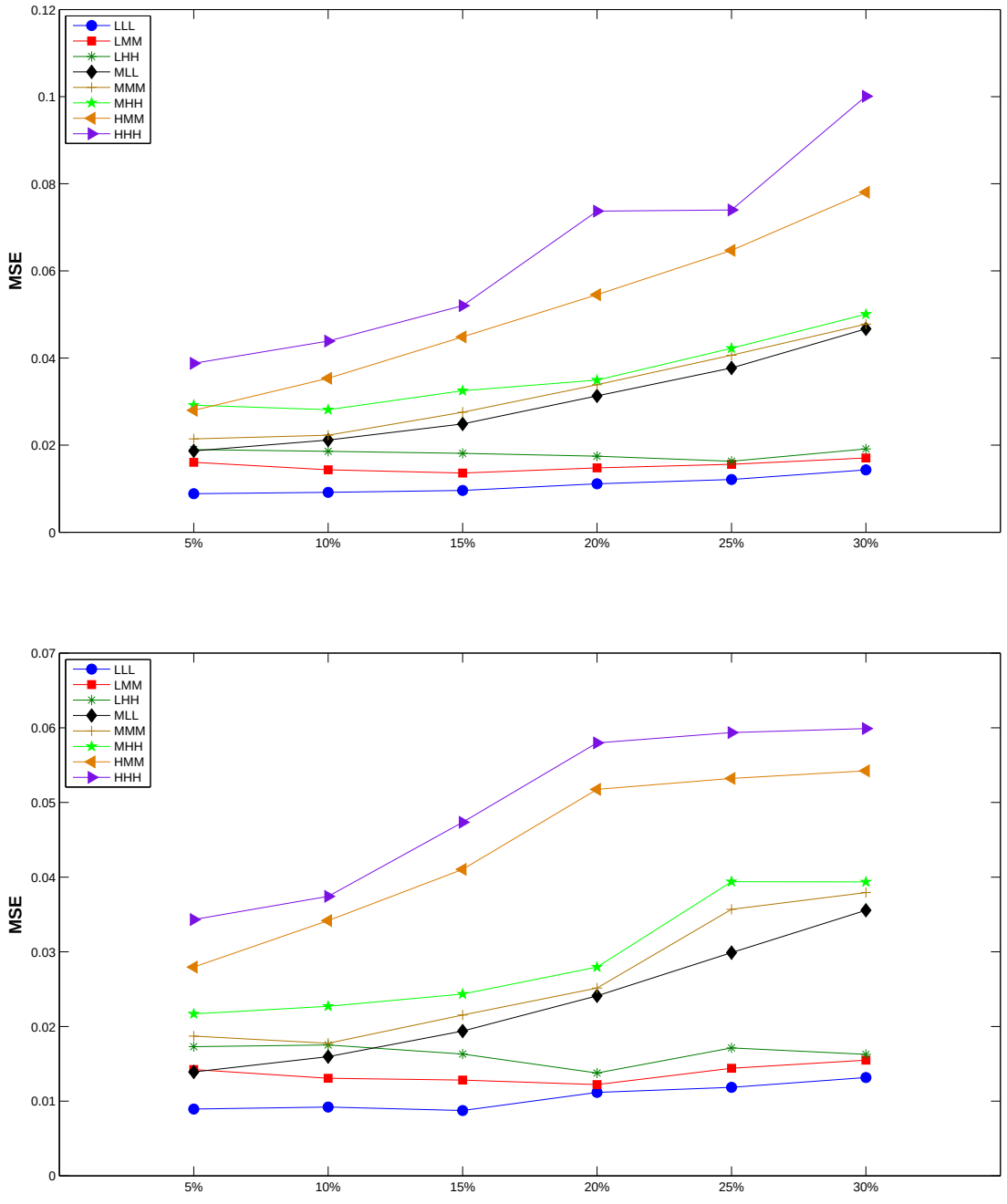


Figure 4.2: Performance of the EM algorithm with increasing percentage of missing values. Upper Panel: Blind-case model; Lower Panel: Informed-case model.

which exemplifies limitations of the popular KNN algorithm. In the figures, the horizonal axis represents data quality and vertical axis represents MSE ratios. From the first block in Fig. 4.5, it is clear that EM outperforms KNN for all qualities of data under consideration. With 10% missing values, although the performance of KNN increases, in 50% cases EM performs better than KNN, as demonstrated in the second block. Also, we observe an increase in the performance of EM in comparison to Mean and Med. A similar conclusion is carried over to the case of 15% missing values. Thus, we conclude that performance of EM in comparison to Mean and Med increases with increasing percentage of missing values (Fig. 4.6). As simulation problems occur with KNN for more than 15% missing values and EM either outperforms or performs almost equivalently as KNN for up to 10% missing values, EM is a better choice among multivariate models to calculate correlation from data sets with small to large percentage of missing measurements.

In Figs. 4.7-4.8 and Figs. 4.9-4.10, we present the performance of all blind-case and informed-case multivariate and bivariate models in terms of MSE values, for different percentage of missing data. The horizontal axis represents various multivariate and bivariate methods used for comparison and vertical axis denotes the MSE values from each method for different data quality. From the figures, we conclude the following: the overall performance of multivariate models is significantly better than the bivariate models; EM remains to be the best performed method for noisy data regardless of percentage of missing values.

## 4.5 Read-world Data Analysis

### 4.5.1 Data

We performed real-world data analysis to further confirm our claims in the preceding section. We tested the performance of various models on two replicated data sets which we considered in Chapter 3:

- spike-in data from Affymetrix (`http://www.affymetrix.com`),

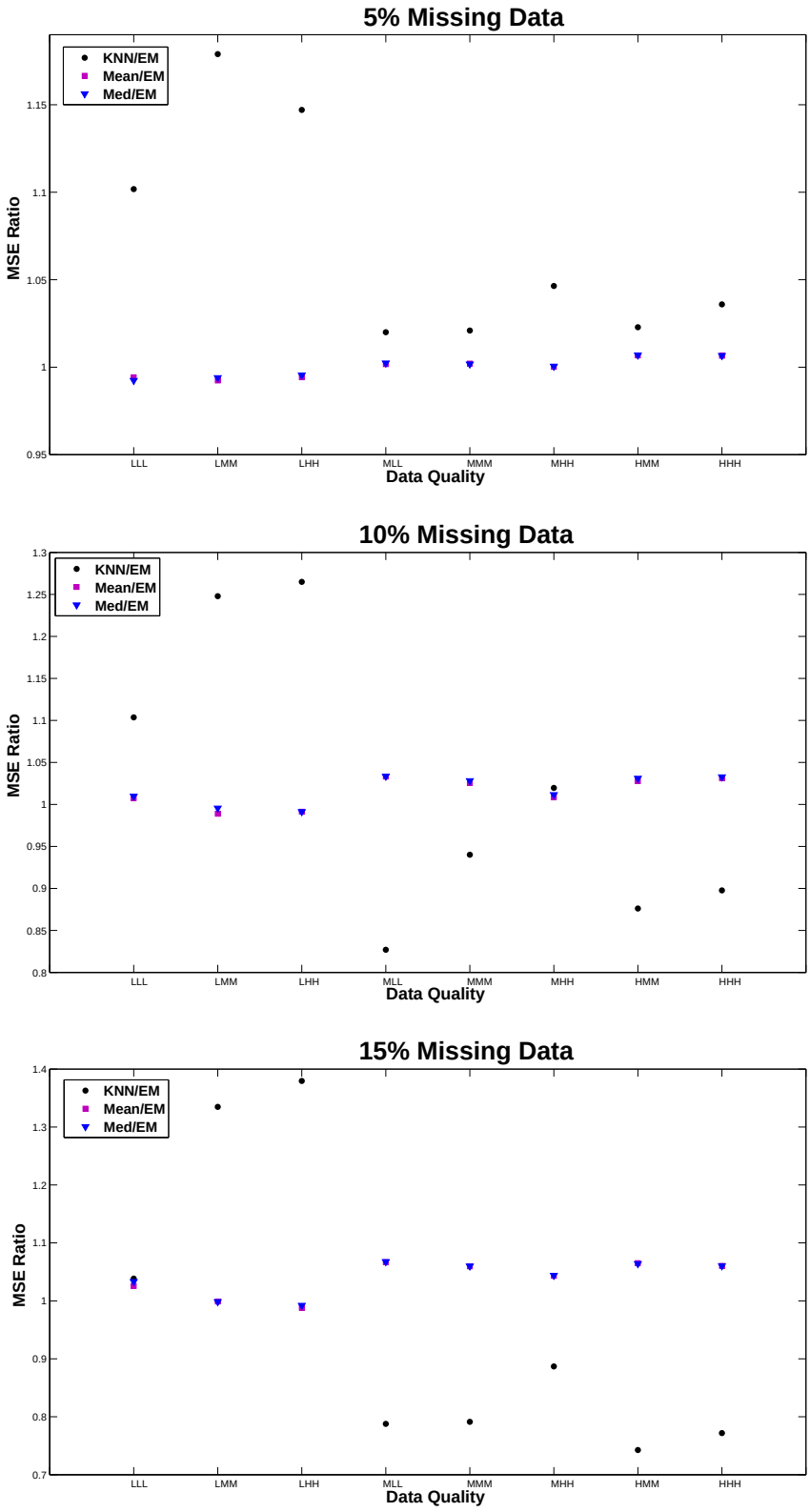


Figure 4.3: Blind-case Model: Comparison of the EM algorithm with other multivariate models, KNN, Mean and Med in terms of MSE ratio (n=20 and m=4). Percentage of missing values is in the range 5%-15%.

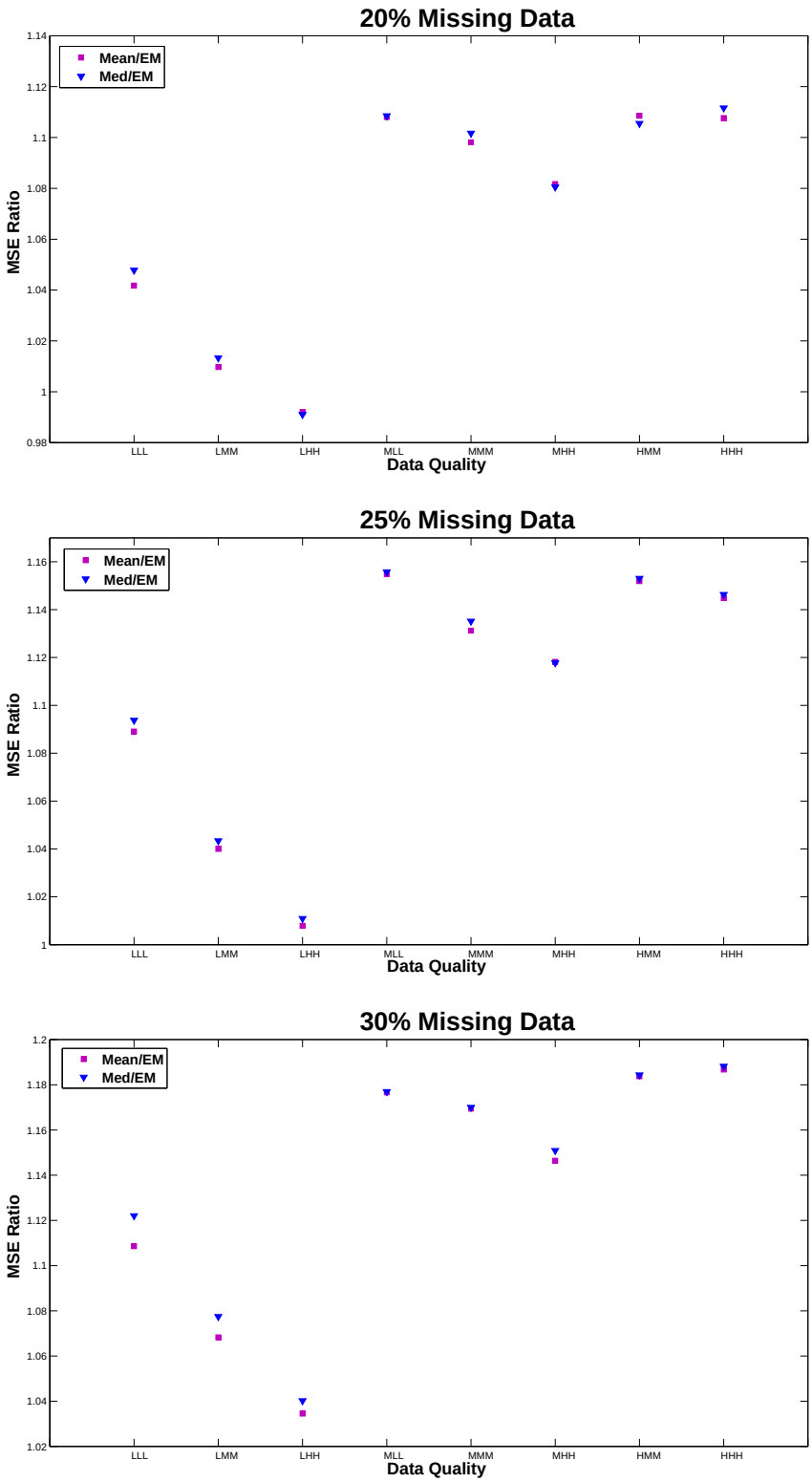


Figure 4.4: Blind-case Model: Comparison of the EM algorithm with other multivariate models, KNN, Mean and Med in terms of MSE ratio (n=20 and m=4). Percentage of missing values is in the range 15%-30%.

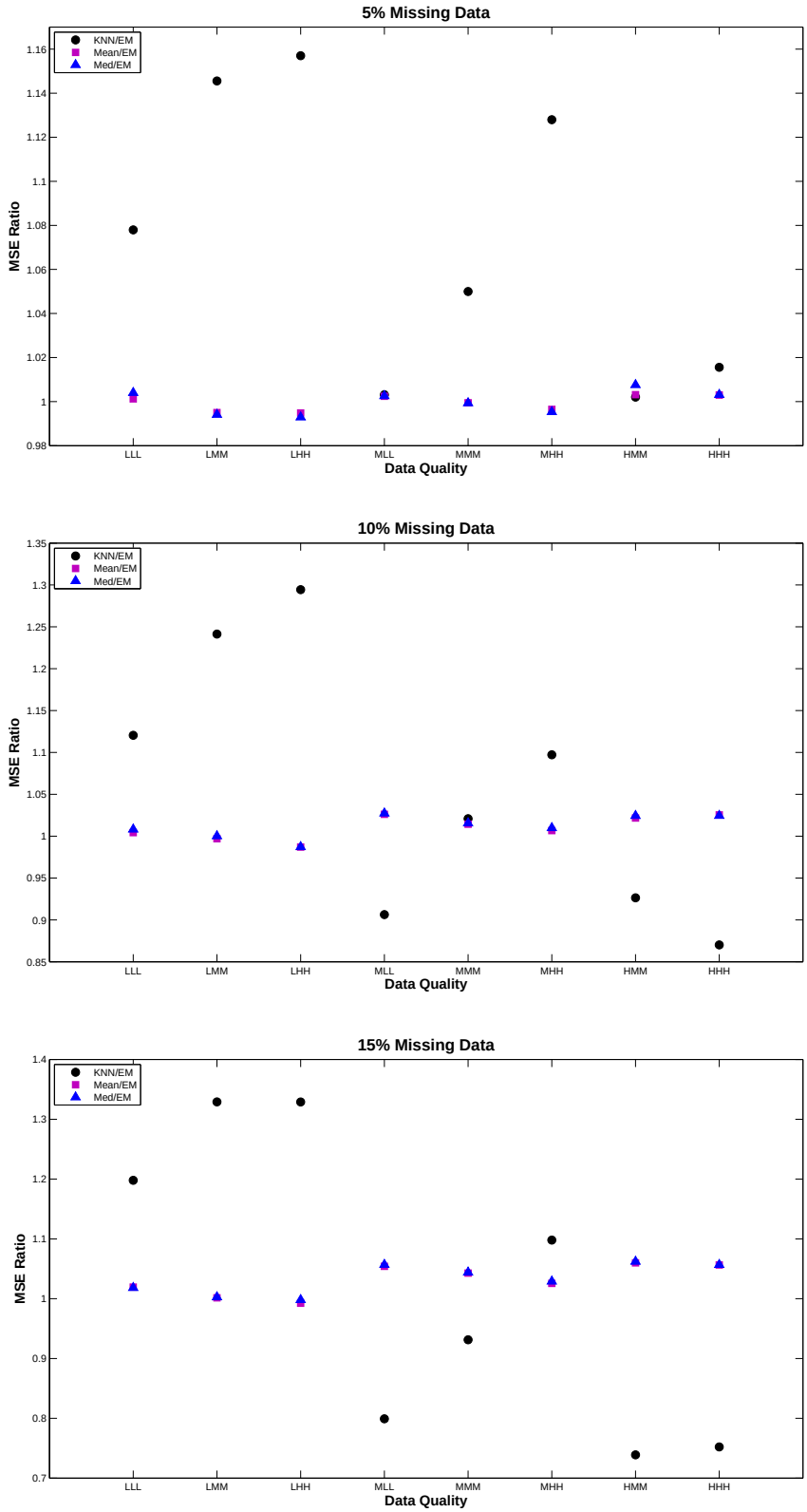


Figure 4.5: Informed-case Model: Comparison of the EM algorithm with other multivariate models, KNN, Mean and Med in terms of MSE ratio (n=20 and m=6). Percentage of missing values is in the range 5%-15%.

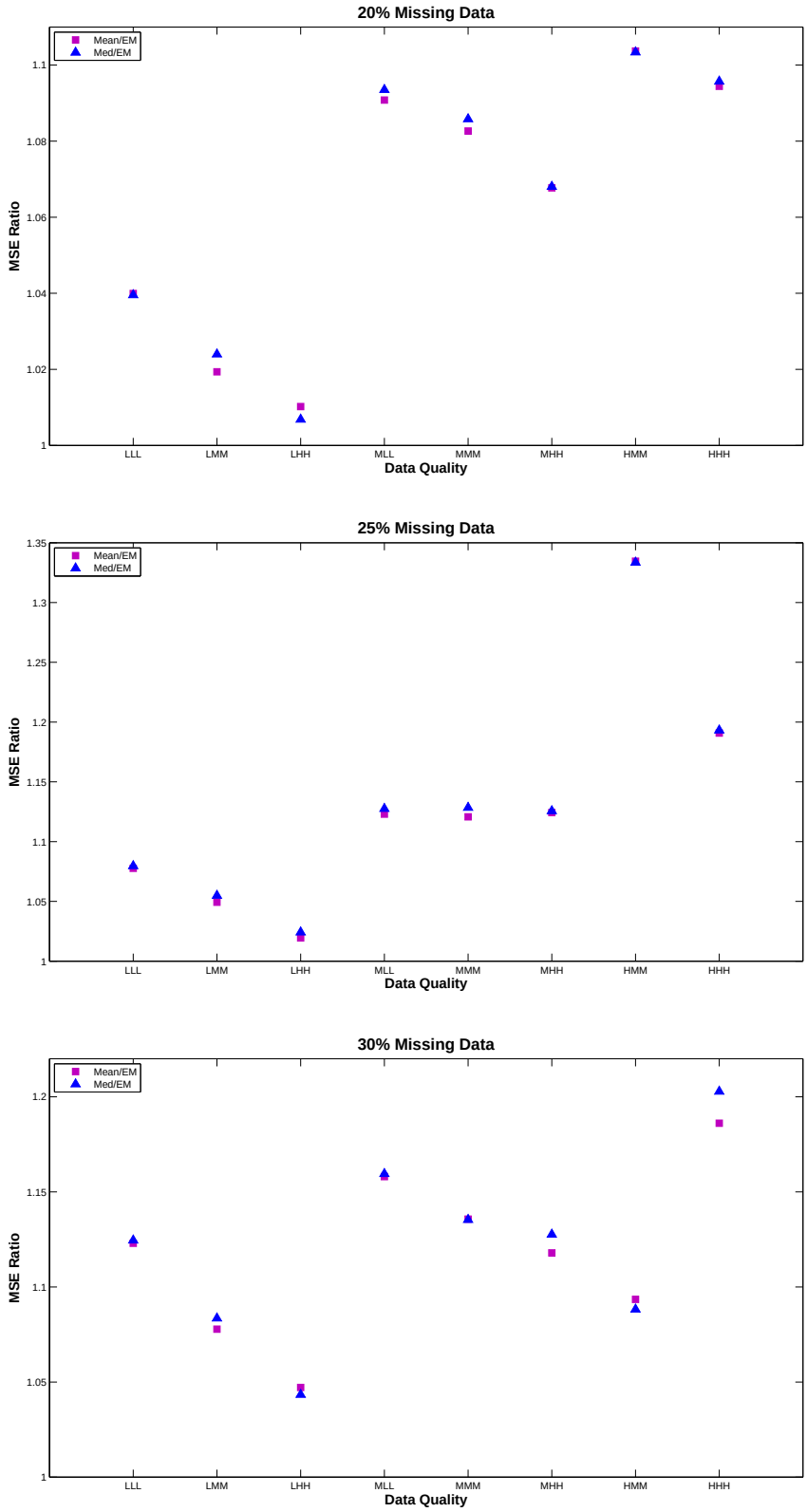


Figure 4.6: Informed-case Model: Comparison of the EM algorithm with other multivariate models, KNN, Mean and Med in terms of MSE ratio (n=20 and m=6). Percentage of missing values is in the range 15%-30%.

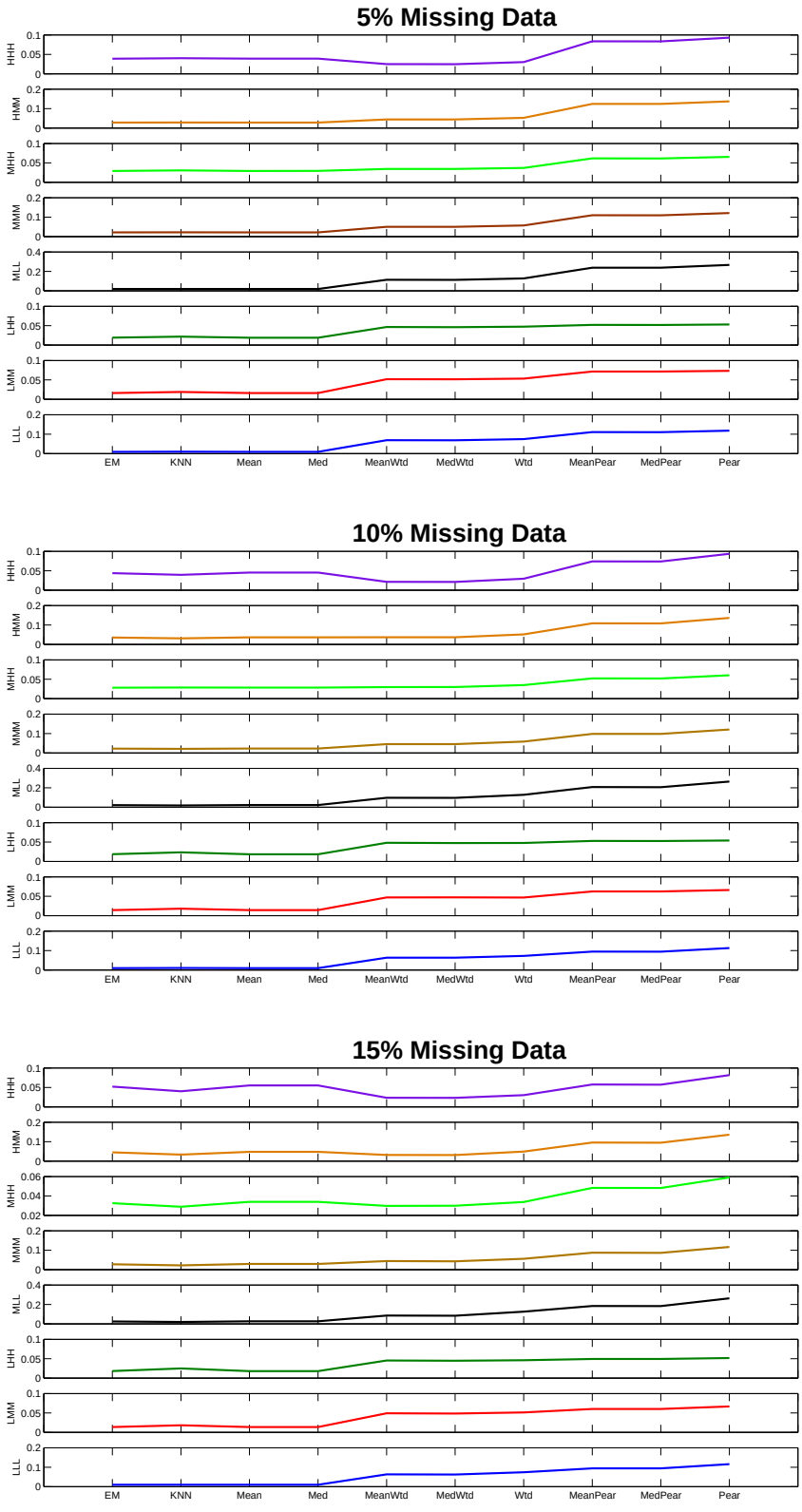


Figure 4.7: Performance of all blind-case multivariate and bivariate models for different percentage of missing values. Percentage of missing value is in the range 5%-15%.

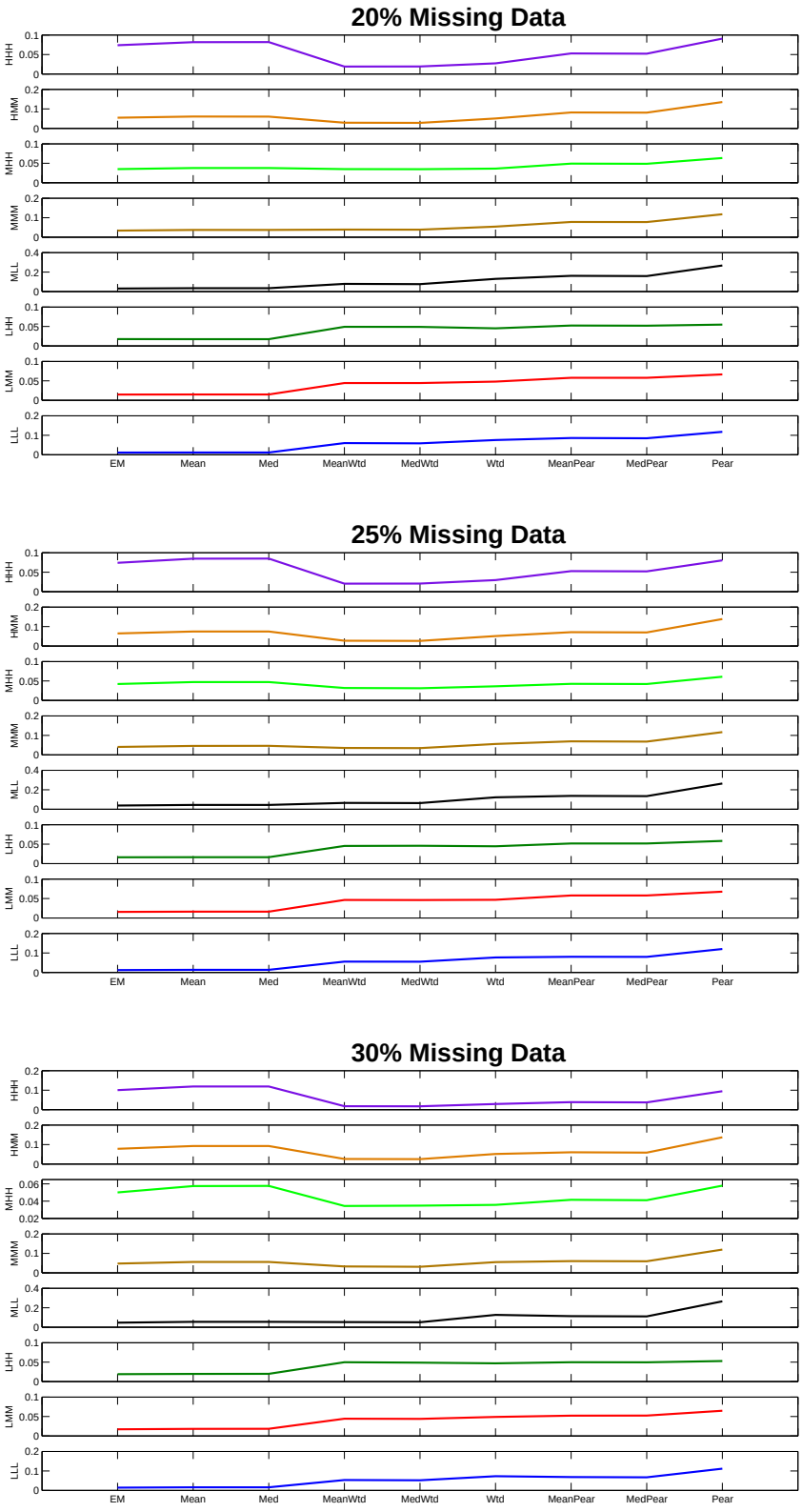


Figure 4.8: Performance of all blind-case multivariate and bivariate models for different percentage of missing values. Percentage of missing value is in the range 15%-30%.

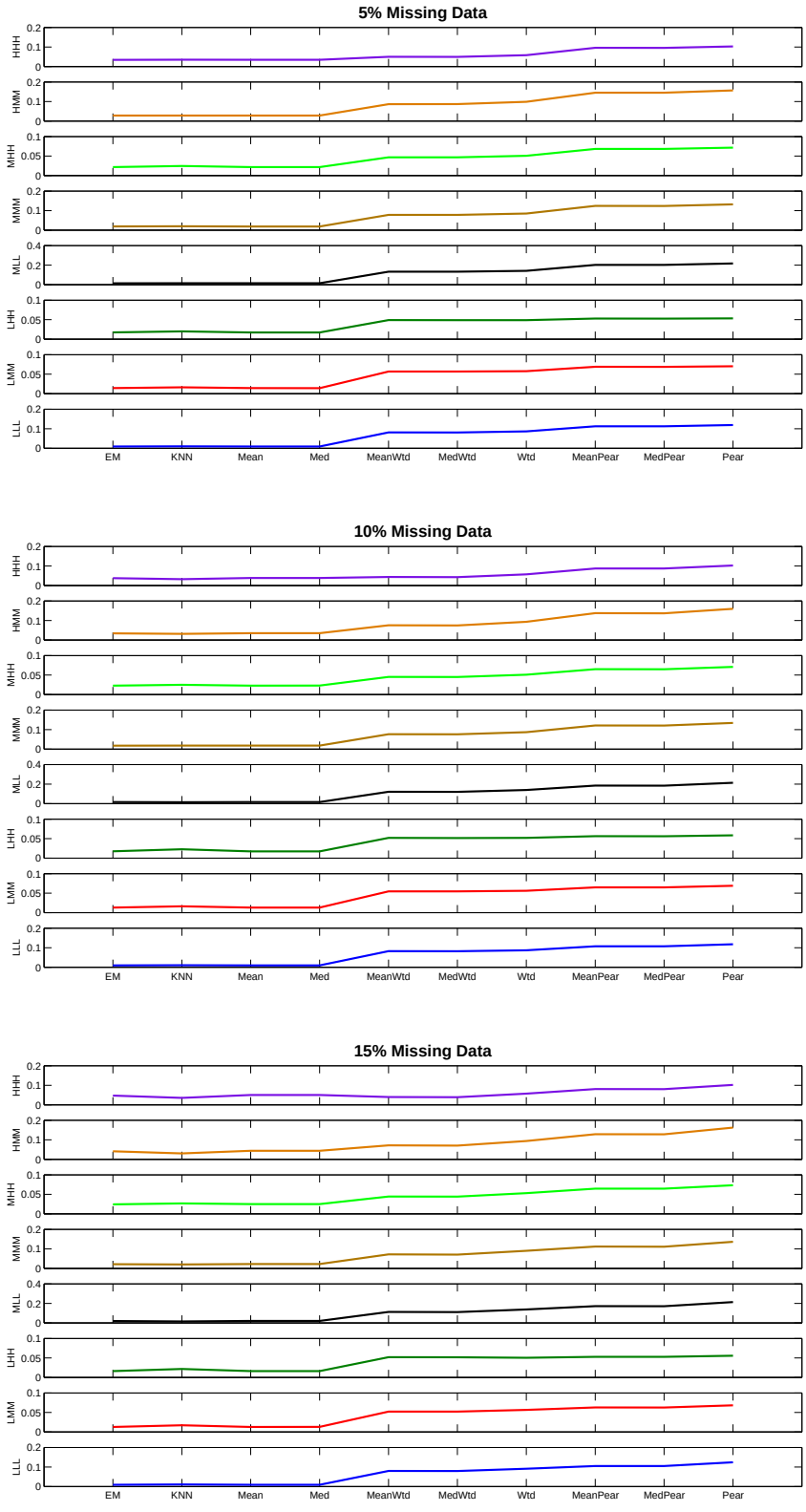


Figure 4.9: Performance of all informed-case multivariate and bivariate models for different percentage of missing values. Percentage of missing value is in the range 5%-15%.

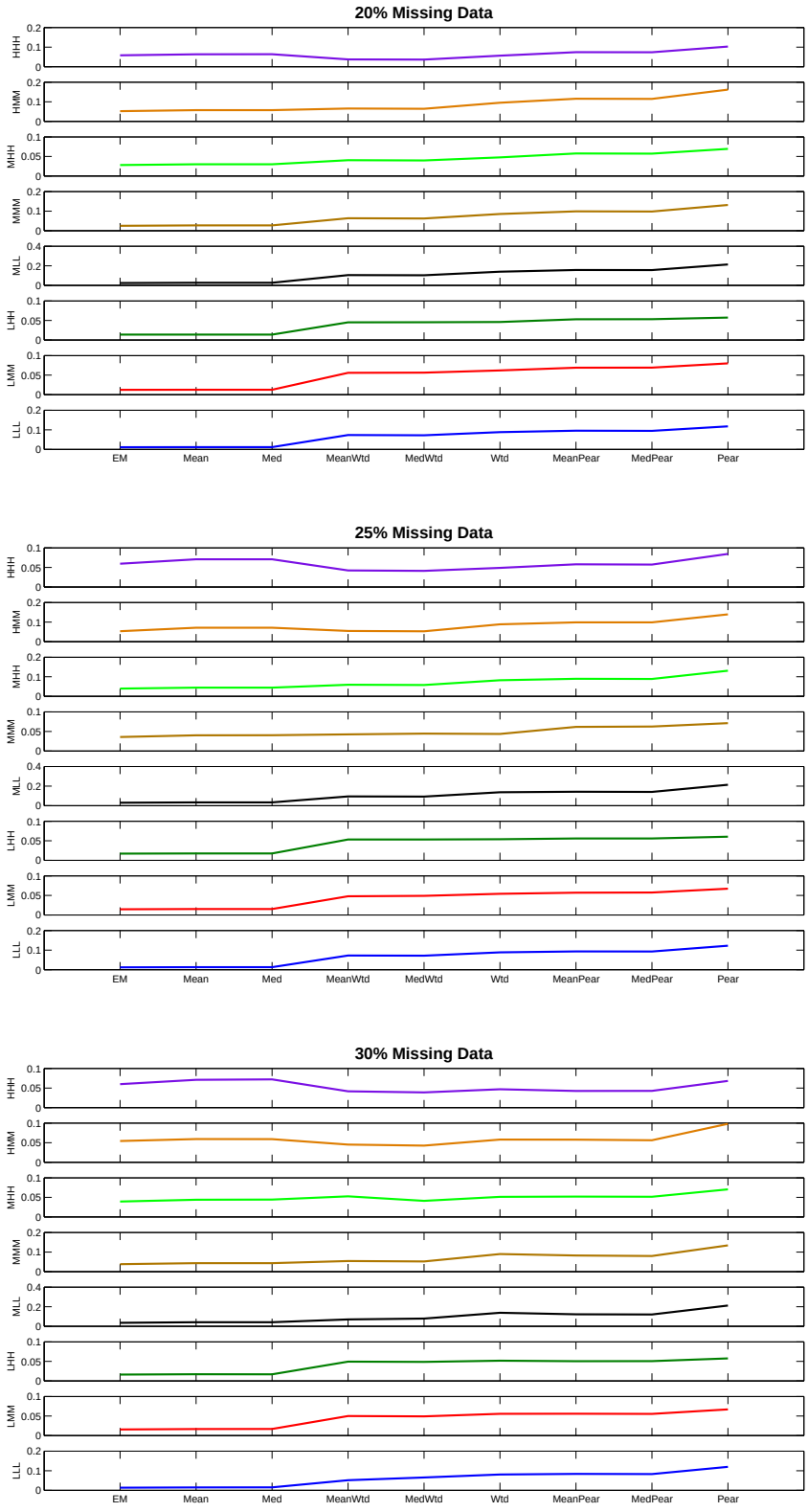


Figure 4.10: Performance of all informed-case multivariate and bivariate models for different percentage of missing values. Percentage of missing value is in the range 15%-30%.

- yeast data (`http://expression.washington.edu/publications/kayee/yeunggb2003`).

We used Affymetrix spike-in data set as benchmark, for which the nominal correlation structure is available, to compare the performances of different models in estimating correlation. The correlation matrix $\hat{\Sigma}$ estimated using different approaches is compared with the nominal correlation matrix $\Sigma$ in terms of MSE values defined as

$$MSE_{\Sigma} = \frac{1}{m^2} \sum_{i=1}^{2m} \sum_{j=1}^{2m} (\Sigma_{ij} - \hat{\Sigma}_{ij})^2.$$

Correlation matrix from a model possessing smaller MSE is closer to the nominal correlation matrix. We randomly removed $5\% - 30\%$ values from spike-in data to compare correlation estimates.

Yeast data set [60], which contains about 8% missing values, was used to test the clustering performance of models. In Chapter 3, we used an imputed version of this data. From the yeast data, we randomly removed data points to make the percentage of missing values vary from 10%, 15%, 20%, 25% to 30%. As mentioned in Chapter 3, the 205 genes in the yeast data were previously classified into four functional groups [152].

### 4.5.2 Estimation of Correlation Structure

In Fig. 4.11, we compared the performances of multivariate models with EM in terms of their MSE ratios, which is the ratio of MSE from a multivariate method over EM. A ratio more than one indicates the better performance of em method. Clearly, from Fig. 4.11, for almost all the cases, MSE ratios are more than one, indicating the superior performance of EM among multivariate models. The real-world data analysis results presented in Fig. 4.11 are also consistent to the ones obtained in simulation studies in that the EM algorithm performs particularly well for replicated molecular profiling with moderate to high percentage of missing values. Fig. 4.12 presents an overall picture of model performances in terms of MSE values. Consistent to what was claimed in our simulation studies, multivariate models

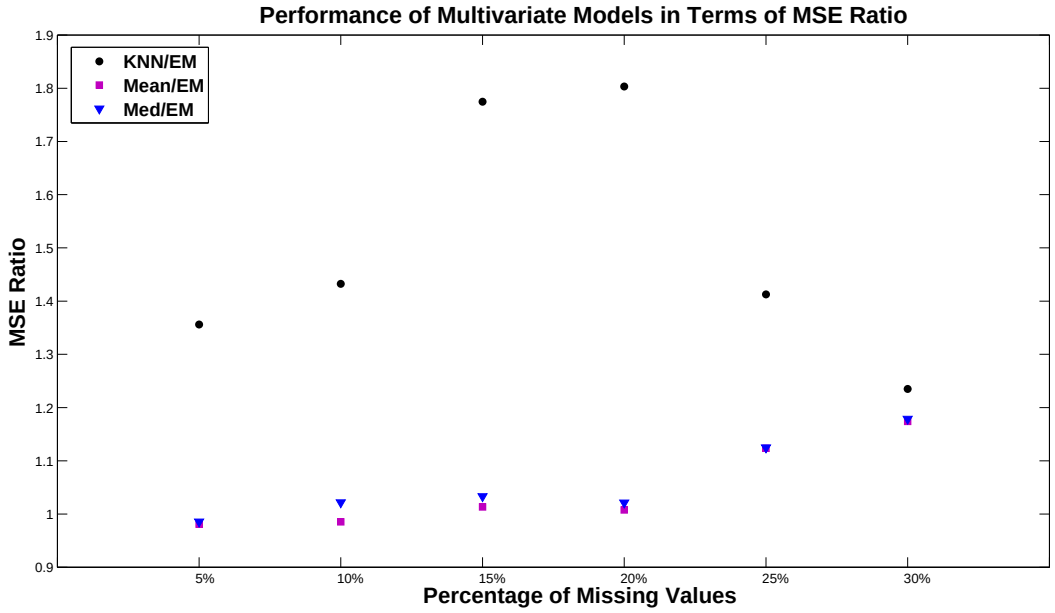


Figure 4.11: Performance of all multivariate models in Affymetrix spike-in data analysis in terms of MSE ratio.

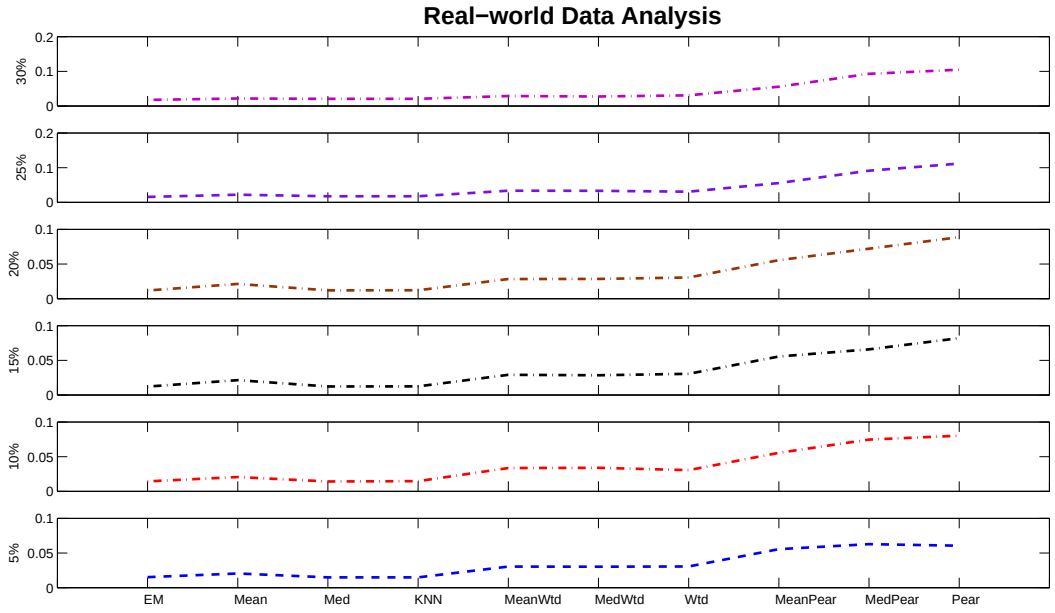


Figure 4.12: Performance of all multivariate and bivariate models in Affymetrix spike-in data analysis for different percentage of missing data.

outperform bivariate ones with EM possessing an overall lower MSE, where we checked the model performances for 5% to 30% missing values.

### 4.5.3 Cluster Analysis

Using yeast data, we calculated a 205 by 205 cophenetic matrix $T$ to represent the external knowledge, where $T(i,j) = 1$ if gene $i$ and gene $j$ are in the same cluster, $T(i,j) = 0$ otherwise. Similarly, for each bivariate and multivariate method, we performed hierarchical clustering by setting the number of the clusters as 4 and calculated a 205 by 205 cophenetic matrix $C^S$. We used these matrices to compute the Minkowski Score defined in Eq. 3.11. Smaller value of the score indicates that clustering is more consistent to the external knowledge. In Figure 4.13, we observed a steady increase in the performance of EM algorithm, reflected by decrease in the Minkowski score, with increasing percentage of missing values. The hierarchical clustering based on EM algorithm performed the best with more than 20% missing values. It further substantiated one of our conclusion that EM algorithm was particularly suitable for calculating correlation matrix from data set with high percentage of missing values.

## 4.6 Discussion

In this chapter, we presented an EM algorithm to estimate the correlation structure from replicated and incomplete molecular profiling data sets. Our approach was based on the assumption that data were random samples from a multivariate Gaussian distribution with a correlation structure given by either blind replication mechanism or informed replication mechanism. Simulation results proved that in most cases the performance of EM algorithm was superior in comparison to other multivariate and bivariate models we considered. In particular, the performance of EM is not sensitive to the data quality when the intermolecular correlation is fixed and the performance of the EM is not sensitive to the percentage of missing values when the intermolecular correlation is low. The intuitive explanation of

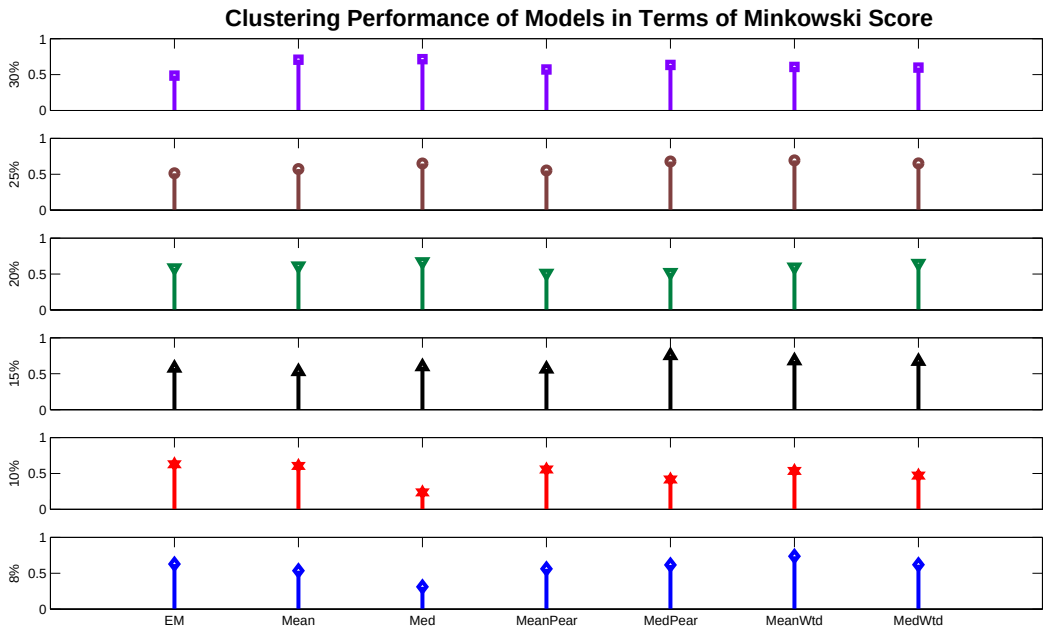


Figure 4.13: Performance of all multivariate and bivariate models in yeast galactose data analysis for different percentage of missing data.

the former is that the MSE is calculated using intermolecular correlation parameters. When they are fixed within a small range, according to Eq. 4.5, the off-diagonal blocks (intermolecular correlation parameters) are not sensitive to variations in the two diagonal blocks (intramolecular correlation parameters, data quality). The intuitive explanation of the latter is low intermolecular correlation between $G_1$ and $G_2$ means random data, and EM treats missing data as random. Therefore, the MSE calculation based on intermolecular correlation parameters is not sensitive to the percentage of missing values.

We also performed real-world data analysis, which confirmed an overall better performance of our approach to the competing ones. Indeed, we tested our approach using Affymetrix spike-in data set by introducing $5\% - 30\%$ missing values. The lower MSE values produced by our approach showed that the correlation values estimated by our approach are closer to the nominal correlation values, which are *a priori* known for the chosen data set. We also used our approach to perform hierarchial clustering with yeast data set. We observed an increase in the clustering performance of EM given by lower Minkowski score,

with increasing percentage of missing values. Thus, our simulation and real-world data analysis results strongly support the use of EM to analyze replicated and incomplete molecular profiling data.

EM was run on a standard desktop computer. The computational complexity of EM was manageable for the chosen set of parameters. In case of incomplete data, EM algorithm utilizing the blind-case model runs faster than the one using informed-case estimator due to lesser number of parameters present in the former estimator. If $k$ is the average number of iterations required by EM algorithm in one simulation, the complexity of blind case estimator is $O(knm)$, where $m$ is the total number of replicated measurements for each biomolecule. The number of iterations in the EM algorithm varies with percentage of missing data, which on an average is $20 - 25$ iterations. Due to the small value of $k$ and the fact that sample size $n$ is kept small in omics experiments for high experimental costs, the algorithm based on blind case estimator is very efficient. The computational complexity of EM algorithm using informed case estimator is $O(knb^2 + kt^2)$, where $b$ and $t$ are the number of biological and technical replicates in each biomolecule, respectively. With the chosen set of parameters, results from EM algorithm in one simulation, using both the estimators could be obtained in less than a minute, however the computational complexity of EM using informed case estimator increases quadratically with increase in the number of biological and technical replicates.

# Chapter 5

# Reconstructing Signaling Pathway Structures: A Sampling-Based Approach

## 5.1 Introduction

In Chapters 2-4, we presented a sequence of multivariate approaches for inferring an optimal correlation structure from high-throughput replicated complete and incomplete molecular profiling data. An accurate estimate of the correlation structure plays a crucial role in various supervised and unsupervised pattern analyses, such as gene clustering [93, 94, 152], inference of gene association networks [19, 20, 125, 126, 155] and gene classification [49, 153], which facilitate the identification of signaling pathway components. In recent years, a wide range of computational tools for analyzing the statistical significance of a gene cluster or a list of differentially expressed genes have become increasingly available due to rapid advancements in high-throughput data acquisition methods [42,56,92,114,140,145]. Although, novel signaling pathway components are revealed by such data analyses techniques, challenges remain in explicitly demonstrating the underlying signal transduction mechanisms in the pathways. In this chapter[1], we specifically focus on the structural inference of signaling pathways, which refers to learning a directed network topology underlying a signaling pathway component.

The structural inference of signaling pathways is important for a better understanding of fundamental cell functions, e.g. growth, metabolism, differentiation and apoptosis [6].

[1]Work accepted for publication. Reused from [2].

Consequently, there have been a wide range of computational efforts for inferring both directed and undirected network topologies. Some of them include Boolean or Probabilistic Boolean networks [66, 135, 136], Bayesian networks [37, 130], ARACNE [88], CLR [35], MRNET [96] and Relevance Networks or RNs [19]. However, the aforementioned approaches primarily focus on statistical causal interactions or pairwise similarities and so the learned networks need not represent signal transduction mechanisms. A few attempts made towards the inference of communication networks from co-occurrence data also find applications in biomedical field [119, 157]. However, significant advantages of exploiting signal transduction mechanisms, which form the basic building blocks of a signaling pathway, are yet to be demonstrated.

We hypothesize a signaling pathway structure as an ensemble of several overlapping signal transduction events with a linear arrangement of genes in each event. We denote these events as Information Flows (IFs). An Information Flow Gene Set (IFGS) contains the genes of the given IF. IFs form the building blocks of a signaling pathway and uniquely determine its structure. The true signaling pathway structure can be reconstructed by inferring the order of genes in each IFGS and combining the inferred IFs into a single unit.

We begin with a compendium of IFGSs related to the pathway. Each IFGS can be interpreted as a discrete set of genes expressed in an experiment, whereas an IFGS compendium comprises of many overlapping IFGSs corresponding to different experiments. Overlapping, which arises from simultaneous participation of genes in many signal transduction events, reflects the interconnectedness among gene sets. We aim to exploit the overlapping among IFGSs to uncover the underlying signal transduction mechanisms.

Since there exist $L!$ different gene ordering permutations for an IFGS with $L$ component genes, the number of signaling pathway structures consistent with a compendium of $m$ IFGSs is of the order of $L!^m$. Neither all network structures are equally likely, nor it is always computationally feasible to find the most likely structure by exhaustive enumeration. In other words, if we treat the ordering of genes in each IFGS as a random variable, which has

a sampling space of size $L!$, it might not be practical to sample directly from the joint distribution of IFGSs with a sampling space of size $L!^m$. As a result, our goal of signaling pathway structure inference can be translated into drawing samples of signaling pathway structures sequentially from the joint distribution of IFGSs and summarizing the most likely structure from the sampled structures. To achieve this goal, we propose a stochastic algorithm, Gene Set Gibbs Sampler (GSGS), developed under the Gibbs sampling framework [40, 41]. GSGS treats the ordering of genes in each IFGS as a random variable, and sequentially samples signaling pathway structures from the joint distribution of IFGSs.

## 5.2 Concepts and Notations

We define an *information flow (IF)* as a directed linear path from one node (usually protein) to another node in a signaling pathway structure, which does not allow self transitions or transitions to a previously visited node. An *information flow gene set (IFGS)* contains the genes of a given IF. Thus, an IFGS and an IF comprise of the same set of genes, however, an IFGS lacks gene ordering information present in the corresponding IF. The length of an IFGS or an IF refers to the number of genes present in it. Clearly, there exist $L!$ different gene ordering permutations for an IFGS of length $L$. Throughout, we assume that $L \geq 3$. IFs of length 2 represent the edges in a signaling pathway structure. We use them to serve as prior knowledge.

Let us consider a compendium of $m$ overlapping IFGSs $X_1,\ X_2, \ldots, X_m$, which we use to infer the underlying signaling pathway structure. We first infer IFs from the IFGSs $X_i$, followed by combining the IFs to represent a single structure. Assuming that the length of $X_i$ is $L_i$, we define a random variable $\Theta_i$ to represent the ordering of genes in $X_i$. The sampling space of $\Theta_i$ is the set of $L_i!$ gene ordering permutations. We write $(X_i, \Theta_i)$ to associate an ordering to the IFGS $X_i$. The notations $\overline{X}$ is used for a given IFGS compendium and we write all IFGSs and their associated orderings together as $(\overline{X}, \overline{\Theta})$, where $\overline{X} = (X_1, \ldots, X_m)$ and $\overline{\Theta} = (\Theta_1, \ldots, \Theta_m)$. The notations are suffixed with $-i$ to consider all but the $i^{th}$ component,

e.g. $\overline{X}_{-i}$, $(\overline{X}, \overline{\Theta})_{-i}$ etc., for $i \in \{1, \ldots, m\}$. We will also utilize the instantiations of $(\overline{X}, \overline{\Theta})$ to construct vectors of size $n \times 1$ and matrices of size $n \times n$, where $n$ is the number of distinct genes among $m$ IFGSs. Suffixing such vectors or matrices with $-i$ means that they have been constructed without involving the $i^{th}$ IF. As the sampling space of $\Theta_i$ is of size $L_i!$, it follows that the sampling space of the joint distribution $P(\overline{X}, \overline{\Theta})$ is the set of $\prod_{i=1}^{m} L_i!$ permutations. A sampling space of size $\prod_{i=1}^{m} L_i!$ can be computationally intractable even for moderate values of $L_i$ and $m$. As a result, our goal of signaling pathway structure inference can be translated into drawing signaling pathway samples sequentially from the joint distribution $P(\overline{X}, \overline{\Theta})$ of IFGSs and summarizing the most likely structure from the sampled pathway structures. Our approach is to develop a Gibbs sampling like algorithm to sequentially sample gene orderings for each IFGS by conditioning on the remaining of the network structures, with a much reduced sampling space of size $L_i!$.

## 5.3 Joint Distribution of IFGSs

We consider IFGSs as random samples from a first order Markov chain model, where the state of a node is only dependent on the state of its previous node. From a given set of $m$ IFs (ordered paths), the two model parameters, initial probability vector $\pi$ and transition probability matrix $\Pi$, are estimated by treating each IF as a Markov chain. If there are $n$ distinct genes across $m$ IFs, we define

$$\pi = (\frac{c_1}{m}, \ldots, \frac{c_n}{m}) \tag{5.1}$$

where $c_l$ is the total number of times $l^{th}$ gene appears as the first node among $m$ IFs, for each $l = 1, \ldots, n$. If $c_{rs}$ is the total number of times $r^{th}$ gene transits to $s^{th}$ gene (i.e. there is edge from $r$ to $s$) among $m$ IFs, then

$$\Pi = [p_{rs}]_{n \times n} \tag{5.2}$$

where $p_{rs} = c_{rs}/\sum_{s=1}^{n} c_{rs},\ r, s = 1, \ldots, n$. Thus, $\Pi$ captures the overlapping signaling mechanisms among IFs.

The parameters $\pi$ and $\Pi$ can be estimated individually for each of the $\prod_{i=1}^{m} L_i!$ collections of IFs. Each collection of IFs is an instantiation of all possible collections and represents a candidate signaling pathway structure. The parameters $\pi$ and $\Pi$ estimated for a collection can be used to calculate its likelihood. The likelihood of a collection of IFs is the product of the likelihoods of $m$ individual IFs in it. The likelihood of each IF can be can be computed by treating it as a first order Markov chain and using the parameters $\pi$ and $\Pi$. For example, we compute the likelihood of the IF $z \rightarrow y \rightarrow x$ as

$$\mathcal{P}(z \rightarrow y \rightarrow x) = P(z) \times P(y|z) \times P(x|y). \tag{5.3}$$

The likelihood values calculated for all $\prod_{i=1}^{m} L_i!$ collections of IFs can be normalized to denote the joint distribution of IFGSs. However, exhaustive computation of $\prod_{i=1}^{m} L_i!$ likelihood values to choose the most likely structure might be computational infeasible, which serves as motivation for the proposed GSGS approach. The computational tractability of GSGS lies in sequentially sampling an order for each IFGS $X_i$ by conditioning on the orders of the remaining IFGSs, with a much reduced sample space of size $L_i!$ as compared to $\prod_{i=1}^{m} L_i!$.

## 5.4 Conditional Distribution of IFGSs

In GSGS, we begin by assigning randomly selected orders to each IFGS. We update the orderings by sampling an order for each IFGS conditioned on the known orders of remaining $m-1$ IFGSs. To sample an order for $X_i$ from the conditional distribution, we leave $X_i$ out. From the remaining $m-1$ IFs, we then compute the initial probability vector $\pi_{-i}$ and transition probability matrix $\Pi_{-i}$ by following the procedure described in Eq. 5.1 and Eq. 5.2. Next, we calculate the likelihoods of all possible orders $\Theta_i^j,\ j = 1, \ldots, L_i!$ for $X_i$ by conditioning on the orders of remaining $m-1$ IFGSs. The normalized conditional likelihood

for the $j^{th}$ order for $X_i$ is given by

$$\mathcal{L}_i^j = \begin{cases} \frac{\mathcal{P}_i^j}{\sum_{j=1}^{L_i!} \mathcal{P}_i^j} & \text{if } \sum_{j=1}^{L_i!} \mathcal{P}_i^j \neq 0, \\ \frac{1}{L_i!} & \text{otherwise} \end{cases} \tag{5.4}$$

where

$$\mathcal{P}_i^j = \mathcal{P}((X_i, \Theta_i = \Theta_i^j)|(\overline{X}, \overline{\Theta})_{-i}), \tag{5.5}$$

where $\mathcal{P}_i^j$ represents the conditional likelihood of the $j^{th}$ order and is computed by decomposing it into the product of conditional probability terms. For example, we compute the conditional likelihood of $z \to y \to x$ corresponding to IFGS $X_i = \{x, y, z\}$ as

$$\mathcal{P}((X_i, \Theta_i = z \to y \to x)|(\overline{X}, \overline{\Theta})_{-i}) = P(z) \times P(y|z) \times P(x|y), \tag{5.6}$$

where each term on the right of Eq. 5.6 is conditioned on $(\overline{X}, \overline{\Theta})_{-i}$ and is available from $\pi_{-i}$ and $\Pi_{-i}$. The $\mathcal{L}_i^j$ values, for $j = 1, \ldots, L_i!$, can now be used to sample an order for $X_i$ from the conditional distribution using inverse Cumulative Density Function (CDF) [40]. The CDF of the conditional distribution $P((X_i, \Theta_i)|(\overline{X}, \overline{\Theta})_{-i})$ is defined as

$$F((X_i, \Theta_i = \Theta_i^j)|(\overline{X}, \overline{\Theta})_{-i})) = \sum_{k=1}^{j} \mathcal{L}_i^k \tag{5.7}$$

for each $j = 1, \ldots, L_i!$. By sampling a number $u \sim U(0, 1)$ and letting $F^{-1}(u) = v$, we get a randomly drawn order $v$ for $X_i$ from the conditional distribution (Eq. 5.7).

## 5.5 Gene Set Gibbs Sampler (GSGS)

In Algorithm 5.1, we present the Gene Set Gibbs Sampling (GSGS) approach, which leads to the reconstruction of signaling pathways from IFGSs. If prior knowledge of some edges (an IF of length 2) is available, we augment them with unordered IFGSs as directed pairs and

**Algorithm 5.1** Gene Set Gibbs Sampler

1: **Input:** $\overline{X} = (X_1, \ldots, X_m)$, where $X_i$'s, $i = 1, \ldots, m$, represent IFGSs, $\overline{E} = (E_1, \ldots, E_e)$, where $E_k$'s, $k = 1, \ldots, e$, represent prior known directed edges (optional), burn-in state $B$ and number of samples $N$ to be collected after burn-in state
2: **Output:** $m$ information flows $(X_i, \hat{\Theta}_i)$, $i = 1, \ldots, m$
3: At $t = 0$, randomly choose an order $\Theta_i^{(0)}$ for $X_i$ from $L_i!$ permutations, $i = 1, \ldots, m$
4: **for** $t = 1, \ldots, B + N$ **do**
5: $\quad \overline{\Theta} = (\Theta_1^{(t-1)}, \ldots, \Theta_m^{(t-1)})$
6: $\quad$ **for** $i = 1, \ldots, m$ **do**
7: $\quad\quad$ Leave $X_i$ out
8: $\quad\quad$ Use the remaining IFs, including those present in $\overline{E}$, to estimate the two Markov chain parameters.
9: $\quad\quad$ Calculate the conditional likelihoods $\mathcal{L}_i^j$'s (Eq. 5.4) of $L_i!$ permutations by treating $X_i$ as a first order Markov chain
10: $\quad\quad$ Sample an order $\Theta_i^{(t)}$ for $X_i$ from the inverse cumulative distribution $F((X_i, \Theta_i)|\overline{E}, (\overline{X}, \overline{\Theta})_{-i})$ (Eq. 5.7)
11: $\quad\quad$ Update the order information for $X_i$
12: $\quad$ **end for**
13: **end for**
14: Return $\hat{\Theta}_i = \text{mode}(\Theta_i^{(B+1)}, \ldots, \Theta_i^{(B+N)})$, $i = 1, \ldots, m$.

keep the direction of genes in each of them fixed during the execution of GSGS. Algorithm 5.1 outputs a list of most frequently occurred IFs among sampled IFs (Step 14 in Algorithm 5.1). To reconstruct a signaling pathway, we start with an empty network of distinct genes present in the input list and infer the most likely signaling pathway by joining IFs present in the output of Algorithm 5.1.

## 5.6 Description of the Case Studies

### 5.6.1 Case Study I: Using the *E. coli* and *In silico* Networks

#### Data

We obtained two gold standard directed networks, *In silico* network [95,139] from DREAM2 and *E. coli* network [86,87,117] from DREAM3 network challenges in the DREAM initiative.

Availability of gold standards allowed us to assess the performance of GSGS using true IFGSs derived from the underlying networks. Both *E. coli* and *In silico* networks comprised of 50 nodes with 62 and 37 true edges, respectively. From the *E. coli* and *In silico* networks, two collections of IFGSs were derived by a direct application of the algorithm presented in Appendix A.7. The algorithm finds unordered IFGSs from a directed network by first finding all IFs (linear paths) in the network and then randomly permuting the order of genes in each IF. There were a total of 125 and 57 IFGSs of length $\geq 3$ for the *E. coli* and *In silico* networks, respectively, which served as input for GSGS. A given percentage of true edges were used to serve as prior knowledge. This study allows us to test: if IFGSs are sampled from the true signaling pathway structure, how well GSGS and other existing approaches can infer the underlying signal transduction mechanisms?

**IFGSs as Binary Discrete Data**

Note that a gene set compendium is essentially a binary discrete data set and *vice versa*. A gene set represents a set of genes expressed in an experiment and so it naturally corresponds to a vector (sample) of binary values obtained by considering the presence (1) or absence (0) of genes in the set. Similarly, genes expressed in a sample of experimental measurements discretized into binary levels, correspond to a gene set. Thus, a gene set and a binary discrete sample represent the same underlying data in two different forms. Keeping this in mind, our approach can be compared with existing network inference approaches accommodating discrete measurements, such as Bayesian networks [26, 37, 100] and mutual information (MI) based network inference methods ARACNE [88], CLR [35], MRNET [96] and RNs [19, 20].

**Comparative Analysis**

In this case study, we compared the performance of GSGS with a number of popular MI based network inference approaches [19, 35, 88, 96] with a primary emphasis on two Bayesian network approaches, K2 [26] and MCMC (Metropolis-Hastings or MH) [100]. The main

reasons are the following:

- From methodology point of view our method infers the most probable linear structure(s) using likelihood scores calculated from the products of conditional probabilities. It is essentially in the same sprit as Bayesian network approaches, while fundamentally different from other approaches which are based on calculating pairwise similarity.

- Both GSGS and Bayesian network approaches take discrete data and infer a directed network. The equivalence between gene sets and binary discrete data makes the comparison between GSGS and Bayesian network approaches very fair.

- Most of the other network inference algorithms, e.g. ARACNE, CLR, MRNET and RNs also discretize continuous data to estimate pairwise similarities, however they are suitable for inferring undirected networks.

A brief description of the aforementioned network inference algorithms has been presented in Appendix A.5 and Appendix A.6, respectively. To compare the performance of GSGS with these approaches, inputs were generated as follows. From the *same underlying network*, e.g. the *E. coli* network,

- We generate IFGSs by a direct application of the algorithm Network2GeneSets [2] presented in Appendix A.7. The IFGSs serve as input for GSGS, whereas the equivalent binary discrete data is used as input for K2, MH and MI based approaches.

- As BN and MI based approaches also accommodate continuous measurements, we generate continuous data inputs for these approaches using Bayes Net Tool Box (BNT) [101, 102].

Performances were evaluated in terms of total number of predicted edges and F-Score (F). A list of performance measures used in this dissertation is as follows:

- Sensitivity (r) = TP/(TP+FN)

- Specificity (s) = TN/(TN+FP)
- Precision or Positive Predictive Value (p) = TP/(TP+FP)
- F-Score (F) = 2pr/(p+r)

where TP = number of true positives, TN = number true negatives, FP = number of false positives, and FN = number of false negatives. An increased number of predicted edges indicate the presence of many false positives, whereas a small number of predicted edges correspond to decreased sensitivity. Total number of predicted edges together with F-Score, reveal an algorithm's performance in predicting true and false positives in a detailed manner.

We used the standard K2 and MH approaches implemented in BNT [101, 102]. The existing implementations were modified to incorporate prior knowledge. Using BNT, we generated continuous measurements from the Gaussian distribution with sample size $20, 30, 40$ and 50. In this case, we used the BIC scoring function. Maximum number of parents allowed for a node in K2 was set at 3. In MH, the burn-in state was set at 500. Number of samples collected after burn-in state was set at 500. For summarizing a network from the sampled ones, an edge present in at least 50% of the networks was declared as true edge. However, we did not observe a significant difference on increasing or decreasing this cut-off. In the case of binary discrete data, both BIC and Bayesian scoring were used. All other parameters were set at default.

We used ARACNE, CLR, MRNET and RN implemented in the R package MINET [97]. In case of continuous data, number of samples was set at 50. Spearman estimator was used to estimate pairwise similarity. In case of discrete data empirical mutual information estimator was used. The MI cut-off applied to an inferred network was set at 0.05. All other parameters were set at default. For example, the DPI parameter `eps` used in ARACNE was set at the default value 0.

### 5.6.2 Case Study II: Using the *E. coli* Data Sets

#### Data

In our second study, we evaluated the performance of GSGS using 4 benchmark *E. coli* data sets available from DREAM3 network challenges in the DREAM initiative [86, 87, 117]. The first two data sets comprise of 50 genes and 51 samples, whereas the remaining two data sets contain 100 genes and 101 samples. The corresponding gold standard networks comprise of 62, 82, 125 and 119 edges, respectively. We derived 4 IFGS compendiums from each of the 4 *E. coli* data sets by declaring the top 10% of the measurements in each data set as 1 and the remaining measurements as 0. This discretization resulted in IFGSs of different lengths across different samples. In each compendium, we considered IFGSs with lengths in the range $3-9$. The resulting compendiums comprised of 47, 45, 45 and 49 IFGSs, respectively.

#### Comparative Analysis

We tested the performance of Bayesian network and MI based methods using *E. coli* data in both, continuous and binary equivalent forms. The performance were measured in terms of F-Scores.

### 5.6.3 Case Study III: Pathway Reconstruction in Breast Cancer Cells

In this study, we analyzed the performance of GSGS by reconstructing a breast cancer signaling pathway from genes present in the ERBB signaling pathway in KEGG database [67, 68]. However, no prior knowledge about the structure of the ERBB signaling pathway available in KEGG was assumed. The ERBB signaling pathway is a directed network of 87 genes and plays an important role in breast cancer signaling [109]. For example, dysregulation/mutation in the epidermal growth factor receptor (EGFR) and ERBB2 (HER2) have been known to promote angiobenesis and metastasis in breast cancer [83, 104].

For network inference, we collected 299 samples of breast cancer patients from Affymetrix

HG-U133 plus 2.0 platform. We mapped all 87 genes participating in the ERBB signaling pathway to the annotation table for Affymetrix HG-U133 plus 2.0 platform, and considered gene expression levels corresponding to exactly one probe set with the highest average measurement among 299 samples for each of them. This resulted in a data set with 87 rows (genes) and 299 columns (samples). IFGSs were derived by discretizing the measurements into binary levels using R package *infotheo*. In our study, we chose the *equalwidth* method to discretize numerical measurements. In a majority of samples ($\sim 66\%$), the number of expressed genes were found in the range $3 - 7$. To compromise between time to reach an appropriate burn-in state and overlapping among gene sets, we considered such samples to form a compendium of 197 IFGSs.

## 5.7 Performance Evaluation

### 5.7.1 Using IFGSs Derived from the *E. coli* and *In silico* Networks

We first analyze the performance of GSGS using the IFGS compendiums derived from the *E. coli* and *In silico* networks. Using GSGS we collected a total of 500 networks after burn-in state which we fixed at 500. All results were averaged from 100 independent runs of GSGS. These parameter were chosen after performing a burn-in state analysis, where we treated sensitivity, specificity and PPV as parameters. A detailed procedure to perform this analysis has been presented in Appendix A.10.

It is worth mentioning here that $P(X_i, \Theta_i)$ may not always be unimodal. A reason which might lead to such a situation is very poor overlapping between $X_i$ and other IFGSs in the compendium. As the discovery of IFGSs depends on the quality of molecular profiling data, it is necessary to test the robustness of GSGS by accommodating real-world under-sampling and over-sampling scenarios. Therefore, we first performed a sensitivity analysis by varying the amount of overlapping among IFGSs. The multi-modal problem is further addressed by incorporating an increasing percentage of prior knowledge and testing if the

algorithm approaches towards the unique true structure.

Fig. 5.1 demonstrates the effect of removing and adding IFGSs to the input of Algorithm 5.1. In Fig. 5.1, $x$-axis represents the percentage of gene sets present in the input, where 20% means that 80% of the gene sets were randomly removed from the original list of IFGSs, and 120% means that 20% of randomly sampled gene sets were added to the list. The figure presents the performance of our approach in terms of the total number of predicted edges. In blocks (a)-(f), the number of edges identified by GSGS (Solid Line) remains close to the ground truth (dashed line). We also observe the positive effect of incorporating prior knowledge. As the percentage of prior knowledge increases (block (a) to block (f)), difference between the ground truth and prediction decreases. In particular, our approach does not produce a large number of false positives in the presence of redundant gene sets.

To further validate the preceding statement, in Table 5.1, we present the F-Scores for the GSGS approach with increasing percentage of gene sets (rows) and prior knowledge (columns). We observe that the F-Scores increase with an increase in the percentage of prior knowledge (values in a row), and these scores remain close on removal or addition of gene sets (values in a column) demonstrating an impressive robustness to under-sampling and over-sampling. This observation strongly supports the applicability of GSGS in the real-world scenarios, where we often do not observe all gene sets or the observed gene sets are redundant.

In Fig. 5.2 and Fig. 5.3, we plot the results from a comparative study in terms of total number of predicted edges using both discrete (left) and continuous (right) data. In the figures dashed line represents the ground truth. It is clear that the number of edges predicted by GSGS remains closer to the ground truth as compared to K2 and MH. In most of the cases, the number of edges predicted by K2 and MH are much higher than the ground truth, indicating an increased number of false positives in the inferred networks.

Fig. 5.4 and Fig. 5.5 plot the F-Scores from different approaches with increasing

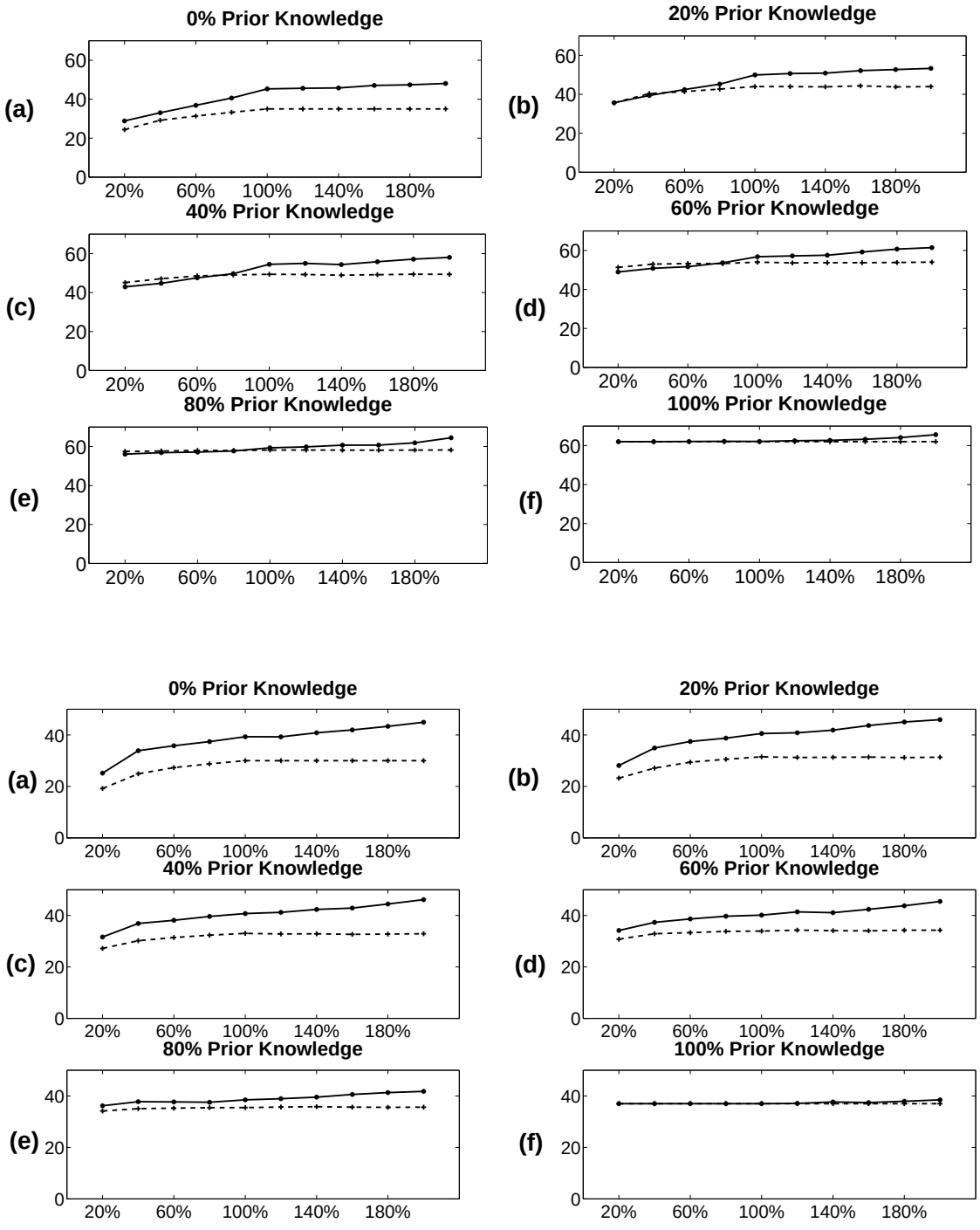


Figure 5.1: Sensitivity analysis for the GSGS approach with increasing percentage of prior knowledge. Network: *E. coli* (Upper Panel) and *In silico* (Lower Panel). In blocks (a)-(f), the $x$-axis represents the percentage of gene sets present in the input and the $y$-axis plots the total number of edges predicted by GSGS (Solid Line). The dashed line plots correspond to the ground truth.

| | 0% | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|
| 20% | 0.430 | 0.648 | 0.748 | 0.844 | 0.926 | 1 |
| 40% | 0.496 | 0.680 | 0.792 | 0.865 | 0.937 | 1 |
| 60% | 0.513 | 0.677 | 0.790 | 0.883 | 0.943 | 1 |
| 80% | 0.468 | 0.665 | 0.780 | 0.860 | 0.947 | 0.999 |
| 100% | 0.457 | 0.595 | 0.719 | 0.824 | 0.923 | 0.999 |
| 120% | 0.459 | 0.590 | 0.704 | 0.825 | 0.913 | 0.996 |
| 140% | 0.450 | 0.579 | 0.722 | 0.805 | 0.909 | 0.999 |
| 160% | 0.422 | 0.564 | 0.691 | 0.803 | 0.913 | 0.991 |
| 180% | 0.434 | 0.550 | 0.679 | 0.786 | 0.897 | 0.984 |
| 200% | 0.425 | 0.546 | 0.676 | 0.778 | 0.877 | 0.974 |

| | 0% | 20% | 40% | 60% | 80% | 100% |
|---|---|---|---|---|---|---|
| 20% | 0.311 | 0.526 | 0.690 | 0.797 | 0.905 | 1 |
| 40% | 0.376 | 0.581 | 0.720 | 0.825 | 0.907 | 1 |
| 60% | 0.448 | 0.596 | 0.737 | 0.818 | 0.918 | 1 |
| 80% | 0.461 | 0.611 | 0.720 | 0.824 | 0.936 | 1 |
| 100% | 0.431 | 0.597 | 0.725 | 0.807 | 0.917 | 1 |
| 120% | 0.448 | 0.591 | 0.715 | 0.790 | 0.913 | 0.999 |
| 140% | 0.412 | 0.555 | 0.686 | 0.788 | 0.900 | 0.992 |
| 160% | 0.414 | 0.539 | 0.661 | 0.762 | 0.884 | 0.995 |
| 180% | 0.403 | 0.499 | 0.644 | 0.745 | 0.867 | 0.989 |
| 200% | 0.372 | 0.497 | 0.612 | 0.717 | 0.858 | 0.982 |

Table 5.1: F-Scores calculated for the GSGS approach with increasing percentage of gene sets in the input (Row) and prior knowledge (Column). Networks: *E. coli* (Left Panel) and *In silico* (Right Panel).

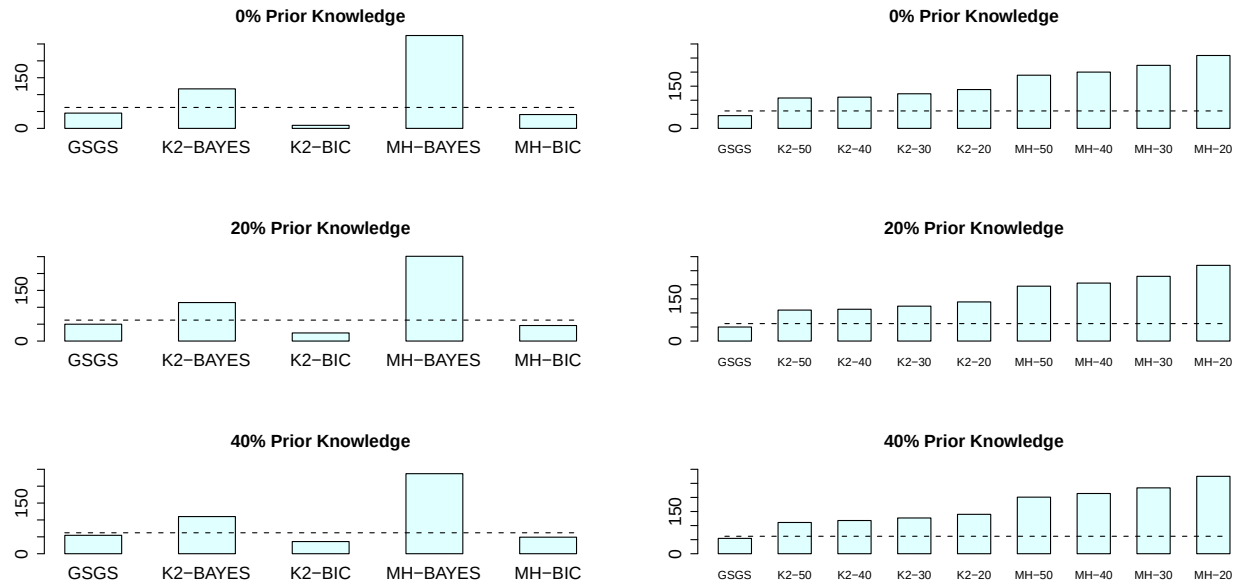


Figure 5.2: Network: *E. coli*. Comparison of the GSGS approach with K2 and MH in terms of the total number of predicted edges with increasing percentage of prior knowledge. Left Panel: Using discrete measurements; Right Panel: Using continuous data with different sample size. The dashed line represents the ground truth.

percentage of prior knowledge. In both the figures, $x$ and $y$ axis represent the percentage of prior knowledge and F-Scores, respectively. We observe that F-Scores for GSGS is significantly higher than K2 and MH using both discrete (upper) and continuous (lower) data. Further, the impact of incorporating prior knowledge on F-Score is more prominent in case of GSGS than K2 and MH, specially on using continuous data where F-Scores for K2 and MH remain much lower than GSGS even in the presence of a large amount of prior knowledge.

We also compared GSGS with four other MI based approaches, ARACNE, CLR, MRNET and RN, without using prior knowledge. The four approaches have been implemented in the R package MINET [97]. As MI networks are undirected, we treated the true underlying networks as well as the networks inferred by GSGS as undirected in the comparison. The F-Scores calculated using both discrete and continuous data are presented in Table 5.2. We observed a significantly better performance of GSGS in comparison to MI network inference methods.

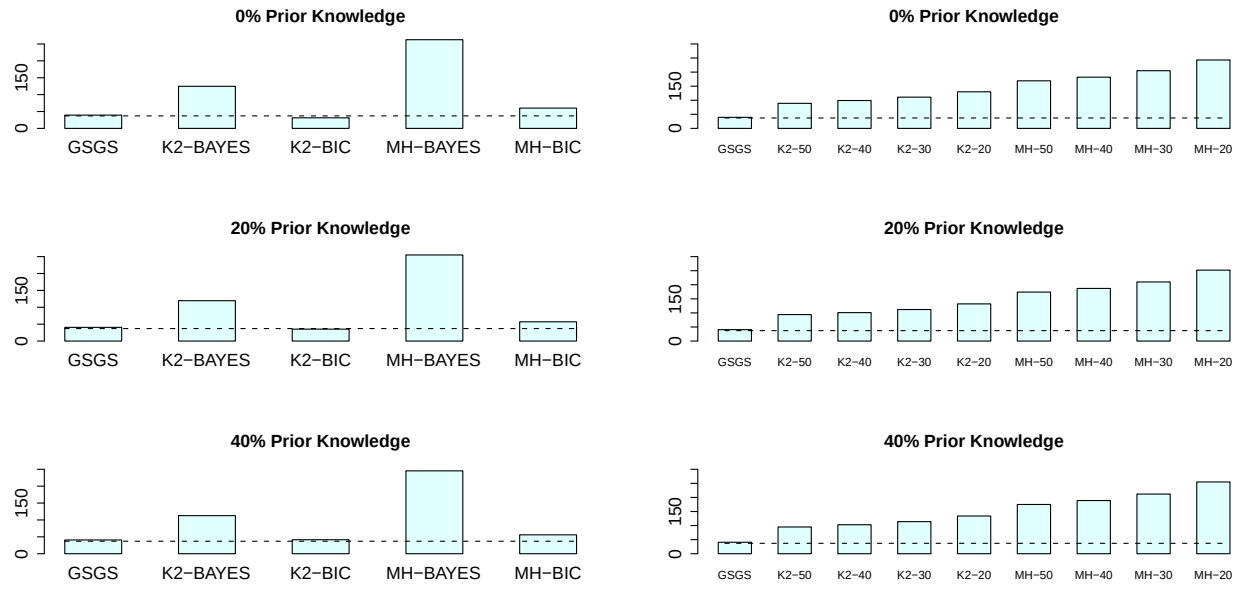


Figure 5.3: Network: *In silico*. Comparison of the GSGS approach with K2 and MH in terms of the total number of predicted edges with increasing percentage of prior knowledge. Left Panel: Using discrete measurements; Right Panel: Using continuous data with different sample size. The dashed line represents the ground truth.

In Figure 5.6, we provide more detailed evidences of the superior performance of GSGS using both *In silico* and *E coli* networks. In Figure 5.6, two left panels represent the true topologies of the two networks, and two right panels represent the topologies reconstructed using GSGS. In each reconstructed network, blue edges represent true positives and gray edges represent false positives. A high level of accuracy is observed in both the reconstructed networks.

### 5.7.2 Using IFGSs Derived from the *E. coli* Data Sets

We applied GSGS to infer signaling mechanisms using the IFGS compendiums derived from *E. coli* data sets. Using each compendium, we collected 500 samples after a burn-in state set fixed at 500. We tested the performance of Bayesian network and MI based methods using the given continuous data sets and binary equivalent data. In each case, we observed a very low sensitivity value by using Bayesian network methods. In addition, we could not

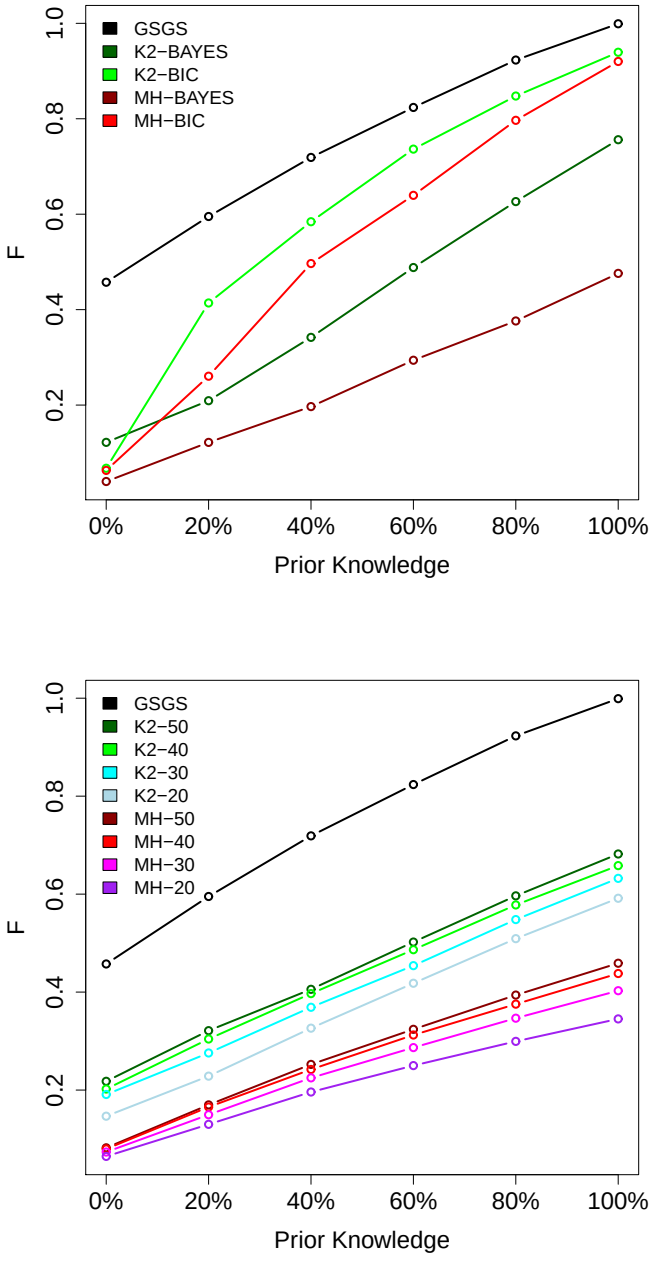


Figure 5.4: Network: *E. coli*. Comparison of the GSGS approach with K2 and MH in terms of F-Scores. Upper Panel: Using discrete measurements; Lower Panel: Using continuous measurements with different sample sizes.

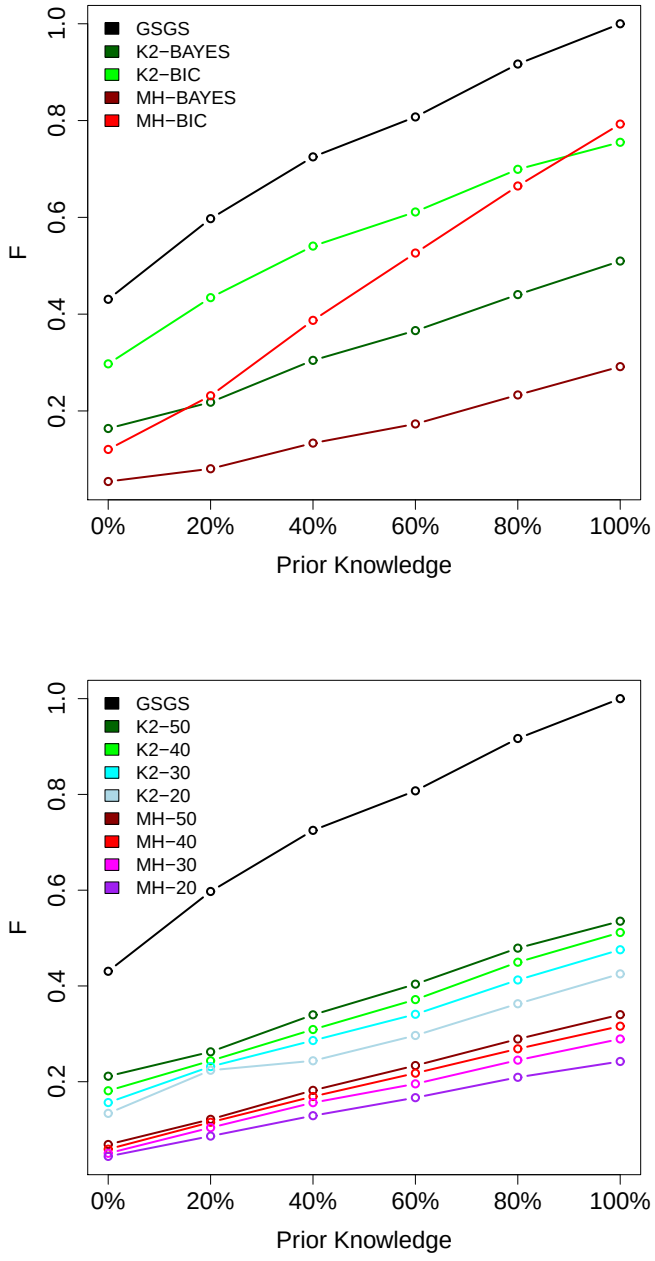


Figure 5.5: Network: *In silico*. Comparison of the GSGS approach with K2 and MH in terms of F-Scores. Upper Panel: Using discrete measurements; Lower Panel: Using continuous measurements with different sample sizes.

| | GSGS | CLR | ARACNE | MRNET | RN |
|---|---|---|---|---|---|
| *E.coli* | 0.79 | 0.40 | 0.18 | 0.18 | 0.25 |
| *In silico* | 0.83 | 0.72 | 0.75 | 0.68 | 0.33 |

| | GSGS | CLR | ARACNE | MRNET | RN |
|---|---|---|---|---|---|
| *E.coli* | 0.79 | 0.39 | 0.54 | 0.47 | 0.30 |
| *In silico* | 0.83 | 0.41 | 0.63 | 0.59 | 0.40 |

Table 5.2: Performance comparison of GSGS with four other pair-wise similarity based network reconstruction approaches in terms of F-Scores. Upper and lower panels correspond to using discrete and continuous data, respectively. For continuous data sample size is 50.

discover any structure in several cases. Therefore, we compared the performance of GSGS with MI based approaches. We inferred MI networks using continuous data as we could not discover a structure in some cases by using discrete data.

In Fig. 5.7, we plot the performance of GSGS and MI based network inference methods in terms of the F-Score ratio, which is the ratio of the F-Score from GSGS and the one from MI based methods. A ratio more than 1 indicates a better performance by GSGS. As shown in Fig. 5.7, we observed a higher F-Score using GSGS, compared with MI based network inference methods.

### 5.7.3 Using IFGSs Related to the ERBB Signaling Pathway

In our final case study, we used GSGS to infer a signaling pathway structure from the IFGS compendium derived using breast cancer molecular profiling data. The IFGS compendium comprised of genes participating in the ERBB signaling pathway in KEGG [67,68]. No prior knowledge about the structure of ERBB signaling pathway available from KEGG was assumed. Using GSGS, we sampled 4500 networks after a burn-in state fixed at 500. The computational complexity of GSGS was easily manageable for the derived IFGS compendium. To validate the performance of GSGS, we utilized the structure of ERBB signaling pathway available from KEGG. As the direction of an information flow is from an upper layer (lower index layer) to a lower one (higher index layer) in the hierarchial representation of a pathway,

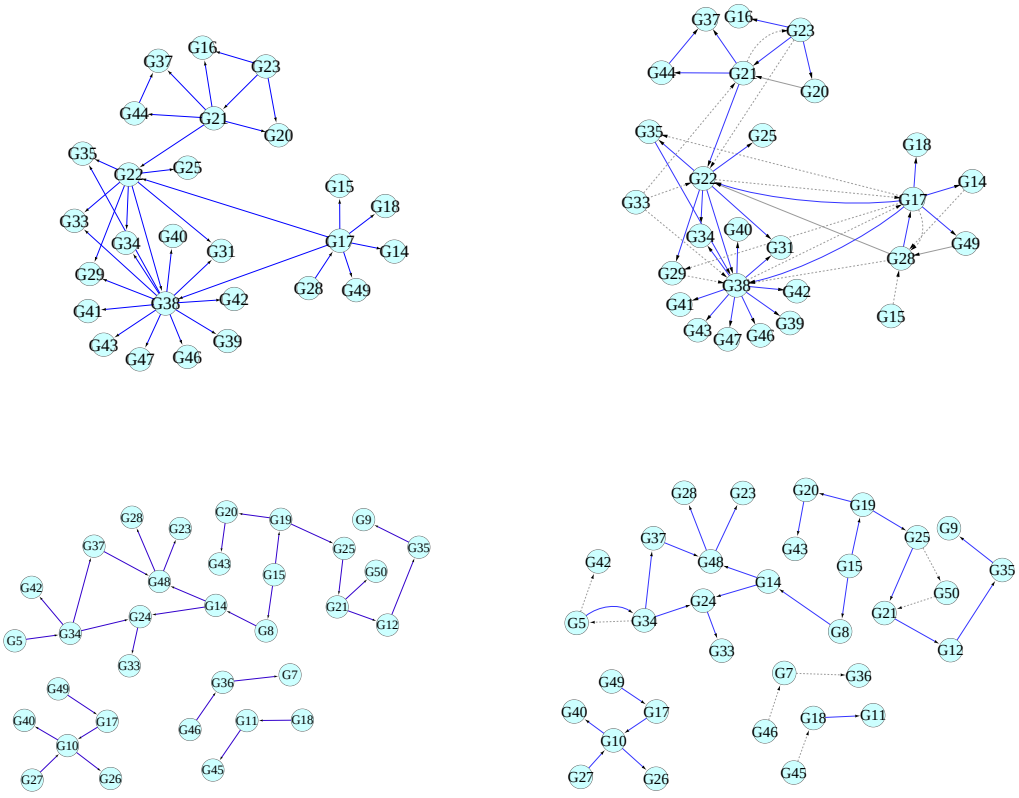


Figure 5.6: A proof of principle study. Left panels show two gold standard networks, *E. coli* (Upper) and *In silico* (Lower). Right panels show the corresponding predicted networks by GSGS, *E. coli* (Upper) and *In silico* (Lower). On the right panels, the blue edges correspond to true positives and gray edges represent false positives. Figures were generated using Cytoscape [131].

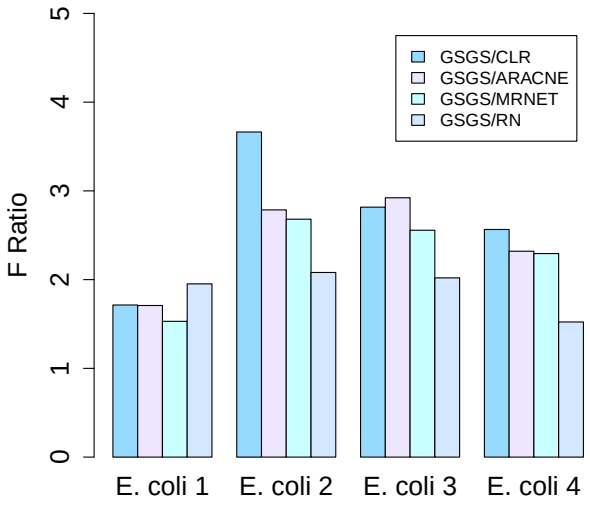


Figure 5.7: Comparison of GSGS with the contemporary MI based network inference methods using four benchmark *E. coli* data sets available from the DREAM initiative.

we collected genes lying in different layers of the ERBB signaling pathway in KEGG, which have been presented in Table 5.3. Considering the noise, i.e. which genes were recognized in each IFGS by data discretization and under-sampling among IFGSs, at the very minimum we expect larger number of edges from a gene in an upper layer to a gene in the lower layer. Indeed, it was found that $\sim 60\%$ of the inferred edges follow this hierarchy, i.e. no parent came from a lower layer. In $\sim 20\%$ of the edges a parent and child node came from the same layer. It is likely that genes lying in the same layer are expressed together in many IFGSs, as they often share a common regulator. Overall, the performance of GSGS depends on purity of input data, like any other inference method.

In the upper panel of Fig. 5.8, we present a few reconstructed signaling events. It can be easily verified that each IF in the figure follows the hierarchy presented in Table 5.3. For example, corresponding to the IFGS {ARAF, ELK1, KRAS}, GSGS predicted an IF KRAS $\rightarrow$ ARAF $\rightarrow$ ELK1, where KRAS came from Layer 6, ARAF from Layer 7 and ELK1 from Layer 10. We further analyzed the inferred structure to identify linear signaling events reported in KEGG. In the lower panel of Fig. 5.8, we present a partial view of the

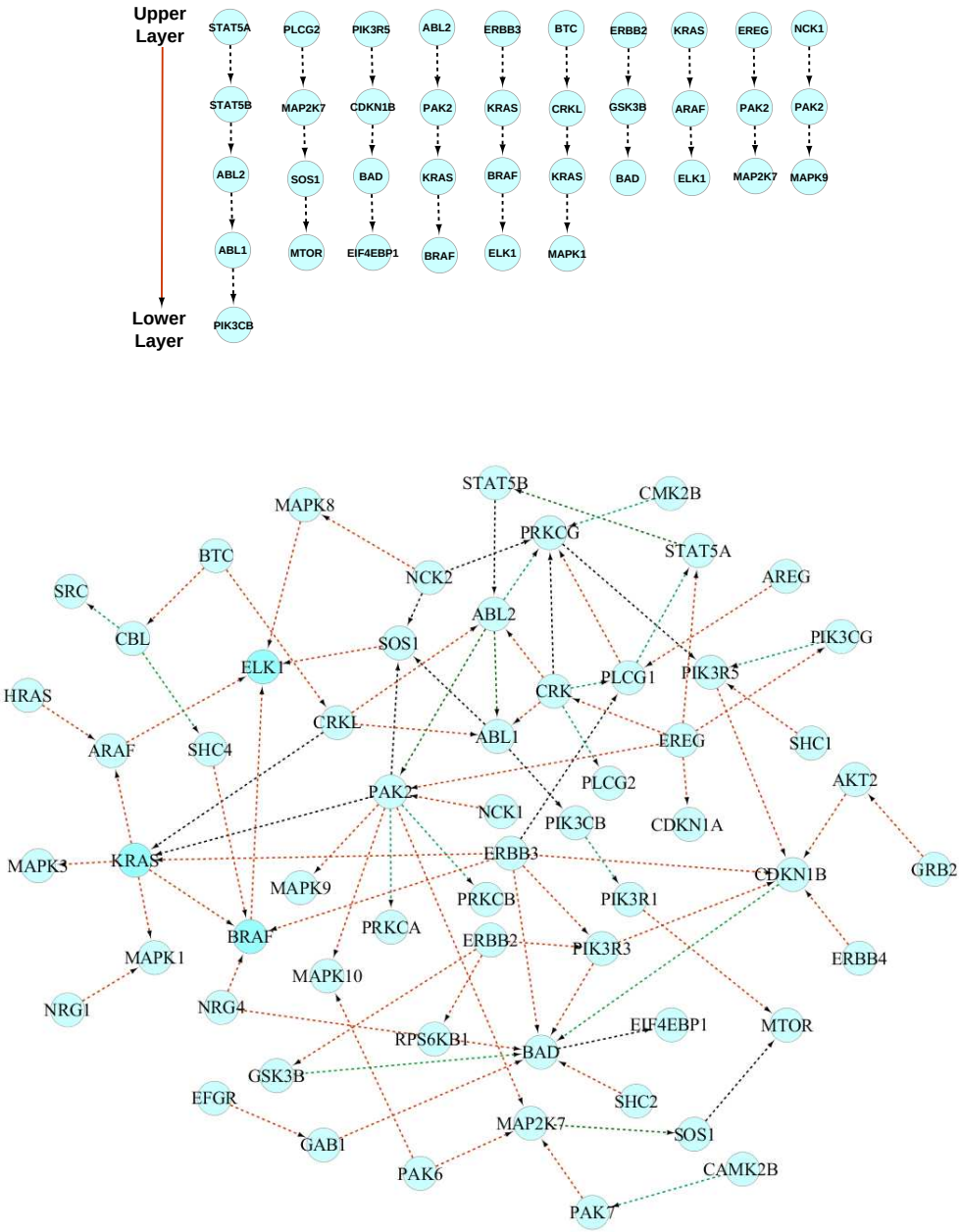


Figure 5.8: Upper Panel: Example of information flows inferred by GSGS. Genes in each information flow follow the hierarchy presented in Table 5.3; Lower Panel: A partial view of the network formed by genes in the neighborhood of ERBB2 and ERBB3. Each information flow follows the hierarchy presented in Table 5.3.

| | Genes |
|---|---|
| Layer1 | EFG, TGFA, AREG, EREG, BTC, HBGEF, ERBB2, NRG1, NRG2, NRG3, NRG4 |
| Layer2 | EGFR, ERBB3, ERBB4 |
| Layer3 | SRC, CBL, CBLB, CBLC, NCK1, NCK2 PLCG1, PLCG2, STAT5A, STAT5B, SHC1,SHC2, SHC3, SHC4, CRK, CRKL |
| Layer4 | PTK2, PAK1, PAK2, PAK3, PAK4, PAK6, PAK7 CAMK2A, CAMK2B, CAMK2D, CAMK2G, PRKCB PRKCA, PRKCG, GRB2, ABL1,ABL2 |
| Layer5 | MAP2K4, MAP2K7, SOS1, SOS2, GAB1 |
| Layer6 | MAPK8, MAPK9, MAPK10, NRAS, HRAS, KRAS, PIK3R2 PIK3CA, PIK3R3, PIK3R5, PIK3CB, PIK3CD, PIK3R1, PIK3CG |
| Layer7 | JUN, ARAF, BRAF, RAF1, AKT1, AKT2, AKT3 |
| Layer8 | MAP2K1, MAP2K2, BAD, MTOR, CDKN1A, CDKN1B, GSK3B |
| Layer9 | MAPK1, MAPK3, E1F4EBP1, RPS6KB1, RPS6KB2 |
| Layer10 | ELK1, MYC |

Table 5.3: Genes arranged in different layers in the hierarchial representation of the ERBB signaling pathway available from the KEGG database.

reconstructed structure formed in the neighborhood of genes ERBB2 and ERBB3. Each edge in the figure follows the hierarchy presented in Table 5.3. Additionally, a red edge means that a linear signaling event between parent and child node has already been recognized in the ERBB signaling pathway structure in KEGG. For example, there exists a linear signal transduction from KRAS to ELK1 via ARAF, and from ERBB3 to ELK1 via KRAS and BRAF in the structure available from KEGG. Green edges correspond to a pair of nodes coming from the same layer in Table 5.3. Black edges represent a pair of nodes, where parent and child node come from an upper and lower layer, respectively, however, a linear signaling between them has not been reported in the pathway structure available from KEGG. Such edges can be viewed as predictions. Overall, the interaction mechanisms presented in Fig. 5.8 support the use of GSGS for inferring signaling pathway structures.

## 5.8 Discussion

In this chapter, we proposed a new sampling based approach, GSGS, to infer the most likely signaling pathway structure from a probability distribution of sampled signaling pathway structures. To achieve our goal, we utilized a compendium of overlapping gene sets related to the given pathway. We first assessed the performance of GSGS by deriving gene sets from two gold standard networks: *E. coli* and *In silico* available from the DREAM initiative. Our approach was shown to have significantly better performance in terms of both F-Score and total number of predicted edges than the Bayesian network approaches K2 and MCMC, and mutual information approaches ARACNE, RN, CLR and MRNET. Robustness of GSGS against under-sampling or over-sampling of gene sets was proved by performing sensitivity analysis. Our conclusions were further validated by testing the performance of the aforementioned approaches on 4 *E. coli* data sets available from DREAM. Finally, we applied GSGS to reconstruct a network in breast cancer cells, and verified it using database knowledge available from KEGG. Overall, our analyses favor the use of GSGS approach in the inference of complicated signaling pathway structures.

As far as we know, GSGS is original in the following aspects: (1). It offers a unique gene set based approach for the reconstruction of directed signaling pathway structures (2). The ordering of genes in each gene set is treated as a random variable to capture the higher order interactions among genes participating in signal transduction events. In most of the existing approaches, individual genes are treated as variables (3). The problem of signaling pathway structure inference is cast into the framework of parameter estimation for a multivariate distribution. (4). The true signaling pathway structures are modeled as a probability distribution of sample signaling pathway structures.

Previously proposed Nested effects models (NEMs) [89] also utilize discrete measurements for inferring a directed structure by constructing it from smaller building blocks. However, a major difference between NEMs and GSGS lies in the fact that NEMs treat

binary effect reporters as random variables, whereas GSGS considers the ordering of genes in IFGSs as random variables. NEM approach builds sub-models by independently scoring all pairs or triplets of genes. An edge in a sub-model is defined in terms of a subset relation between phenotypic profiles of two genes. GSGS, on the other hand, infers gene ordering in a gene set by utilizing the overlapping among all of the remaining gene sets, which more naturally captures the higher order interaction mechanisms. GSGS further benefits from allowing a building block of larger size and explicitly accommodating linear signal transduction mechanisms in its settings, which characterize a signaling pathway structure.

The worst case time complexity of GSGS is $Nm(m+n+ML)$, where $N$ is the number of sampled pathways, $m$ is the number of IFGSs, $n$ is the number of distinct genes, $L$ is the length of the longest gene set in the input and $M = L!$. As longer gene sets ($L \geq 10$) are less likely to correspond to linear information flows, the complexity arising from $ML$ could be managed by appropriately selecting the length of gene sets in each experiment. It is worth mentioning here that GSGS benefits from a much reduced computational load, both in terms of speed and memory requirements, in comparison to Bayesian network approaches, e.g. BN inference using sampling based Metropolis-Hastings approach. Indeed, the complexity of GSGS is driven by the number of possible orderings for IFGSs, which is comparatively much smaller than the number of neighbors of a network generated at each stage of Metropolis-Hastings approach. Complexity of Metropolis-Hastings approach is often unmanageable even for a network of small size as a large number of neighboring structures need to be stored for sampling the next structure.

# Chapter 6

# Reconstructing Signaling Pathway Structures: A Discrete Optimization Approach

## 6.1 Introduction

In the previous chapter, we presented a sampling algorithm to infer signaling pathway structures from gene sets related to the pathways. Since we hypothesized gene sets as discrete measurements emitted from latent signaling pathway structures, another way to approach this problem of combinatorial nature is to utilize discrete optimization based search techniques. In many practical situations, search algorithms rather than sampling based approaches may be advantageous for a variety of reasons, such as simple computational framework, easy implementation and reduced computational burdens. In this chapter[1], we utilize the framework of simulated annealing [74], a widely used search algorithm for addressing discrete optimization problems.

Simulated annealing or SA [74] is a global search algorithm and has its root in the field of metallurgy, where a metal is heated and then cooled down slowly so that the atoms gradually configure themselves in states of lower internal energy, refining the crystalline structure of the metal. Compared with other global search algorithms such as genetic algorithm [53] and tabu search [43], SA is easier to understand and to implement without sacrificing performance. Since genetic algorithm is a population based search method and tabu search is a memory based heuristic, each iteration of SA runs faster than the two approaches. SA also

[1]Work under review. Reused from [3].

requires a small number of user-specified parameters. In the past, SA has inspired various bioinformatics researches [13, 23, 45].

Similar to Chapter 5, we utilize a compendium of IFGSs related to a signaling pathway component and propose a gene set based simulated annealing algorithm, GSSA, by treating IFGSs as the basic building blocks of a signaling pathway. The proposed algorithm mimics the physical process of heating and then cooling down a substance slowly to obtain a strong crystalline structure, by annealing gene sets to infer signaling cascades characterizing the optimal signaling pathway structure. Throughout, we treat unordered gene sets as random variables and their orders as random.

## 6.2 Notations

We denote an information flow gene set (IFGS) introduced in Chapter 5 by $X_i$ and an information flow (IF) by $(X_i, \Theta_i)$, where $\Theta_i$ represents an instantiation of gene orderings in $X_i$, $i = 1, \ldots, m$. We denote an IFGS compendium and a signaling pathway structure by $\overline{X}$ and $(\overline{X}, \overline{\Theta})$, respectively, where $\overline{X} = (X_1, \ldots, X_m)$ and $\overline{\Theta} = (\Theta_1, \ldots, \Theta_m)$. We construct a signaling pathway structure $(\overline{X}, \overline{\Theta})$ by combining the IFs $(X_i, \Theta_i)$ into a single unit. The length of an IFGS $X_i$, which is the number of genes present in $X_i$, is denoted by $L_i$.

## 6.3 A Discrete Optimization Problem

Since there exist $L_i!$ different gene ordering permutations for the IFGS $X_i$, a total of $\prod_{i=1}^{m} L_i!$ distinct signaling pathway structures can be constructed from $\overline{X}$. To locate the true signaling pathway structure, we formulate the following discrete optimization problem

$$\min_{(\overline{X},\overline{\Theta}) \in \mathcal{F}_{\overline{X}}} \mathcal{E}(\overline{X}, \overline{\Theta}) \tag{6.1}$$

where $\mathcal{E}(\overline{X}, \overline{\Theta})$ is the *energy* of the structure $(\overline{X}, \overline{\Theta})$ and $\mathcal{F}_{\overline{X}}$ is the *feasible set* containing the set of candidate signaling pathway structures. Thus, the true signaling pathway structure can

be inferred by (1) defining the energy $\mathcal{E}(\overline{X}, \overline{\Theta})$ (2) defining the feasible set $\mathcal{F}_{\overline{X}}$ of candidate signaling pathway structures such that the true structure has the lowest energy among the candidates and (3) searching for the true signaling pathway structure in $\mathcal{F}_{\overline{X}}$.

## 6.4 Energy of a Signaling Pathway Structure

We propose a novel function to score a candidate signaling pathway structure by treating IFGSs as random samples from a first order Markov chain model. The score of a signaling pathway structure $(\overline{X}, \overline{\Theta})$ is interpreted as its energy and is defined as

$$\mathcal{E}(\overline{X}, \overline{\Theta}) \;=\; -\sum_{i=1}^{m} \log \ell(X_i, \Theta_i), \tag{6.2}$$

where $\ell(X_i, \Theta_i)$ stands for the likelihood of IF $(X_i, \Theta_i)$. Indeed, we compute the likelihood of $(\overline{X}, \overline{\Theta})$ as

$$\mathcal{L}(\overline{X}, \overline{\Theta}) = \prod_{i=1}^{m} \ell(X_i, \Theta_i). \tag{6.3}$$

Since log function is monotonically increasing, searching for a structure with the maximum likelihood is equivalent to seeking a structure with the minimum energy. Each likelihood term $\ell(X_i, \Theta_i)$ is computed using the estimates of two Markov chain parameters, the initial probability vector $\pi$ and the transition probability matrix $\Pi$, which we defined in Section 5.3. Recalling from Section 5.3, the likelihood of an IF, say $x \to y \to z$ is computed as

$$\ell(x \to y \to z) = P(x) \times P(y|x) \times P(z|y), \tag{6.4}$$

where prior and conditional probability terms in the above equation are known from $\pi$ and $\Pi$. The energy of a structure $(\overline{X}, \overline{\Theta})$ can now be computed using Eq. 6.2.

**Algorithm 6.1** Gene Set Simulated Annealing

**Input:** IFGSs $X_i,\ i = 1, \ldots, m$, cooling schedule constant $c$, number of jumps $J$.
**Output:** The reconstructed signaling pathway structure.
**Initialization:** At $k = 0$, randomly select a feasible structure $(\overline{X}, \overline{\Theta}^{(0)})$. Let BestNetwork $= (\overline{X}, \overline{\Theta}^{(0)})$ and BestEnergy $= \mathcal{E}(\overline{X}, \overline{\Theta}^{(0)})$.
**for** $k = 1, \ldots, J$ **do**
  Randomly choose a network $(\overline{X}, \overline{\Phi})$ from the neighborhood of $(\overline{X}, \overline{\Theta}^{(k-1)})$, where $\overline{\Phi} = (\Phi_1, \ldots, \Phi_m)^T$.
  **if** $\mathcal{E}(\overline{X}, \overline{\Phi}) < \mathcal{E}(\overline{X}, \overline{\Theta}^{(k-1)})$ **then**
    $\overline{\Theta}^{(k)} = \overline{\Phi}$
    **if** $\mathcal{E}(\overline{X}, \overline{\Phi}) <$ BestEnergy **then**
      BestNetwork $= (\overline{X}, \overline{\Phi})$
      BestEnergy $= \mathcal{E}(\overline{X}, \overline{\Phi})$
    **end if**
  **else**
    Draw a Bernoulli sample with probability of TRUE as $\min\{1, \exp(\mathcal{E}(\overline{X}, \overline{\Theta}^{(k-1)}) - \mathcal{E}(\overline{X}, \overline{\Phi})/T_k)\}$.
    **if** TRUE **then**
      $\overline{\Theta}^{(k)} = \overline{\Phi}$
    **end if**
  **end if**
**end for**
Return BestNetwork.

## 6.5 Feasible Signaling Pathway Structures

Not all $\prod_{i=1}^{m} L_i!$ signaling pathway structures, which can be constructed from $\overline{X}$, exhibit the topological properties of real-world biological networks. To eliminate random structures from the search space, we only consider candidates which possess certain low-level topological properties such as the degree distribution of underlying structure. The degree distribution of underlying signaling pathway structure, say $(\overline{X}, \overline{\Theta})$, is a weighted asymmetric adjacency matrix $W$ obtained by counting the number of occurrences of directed edges between all gene pairs among $m$ IFs $(X_i, \Theta_i)$, $i = 1, \ldots, m$. Note that except for the pair of terminal nodes, the incoming and outgoing degrees of all intermediate nodes in an IF is 1. Since we consider $(\overline{X}, \overline{\Theta})$ as a set of information flows, it can be easily verified that structures obtained by randomly permuting the orders of intermediate nodes in each IF $(X_i, \Theta_i)$, $i = 1, \ldots, m$, also have degree distribution $W$. Such structures preserve the marginal degree distributions of

genes and form the feasible set $\mathcal{F}_{\overline{X}}$ of size $\prod_{i=1}^{m}(L_i-2)!$. In simulation studies, $W$ can be obtained from the true signaling cascades. In real-world studies, it can be approximated by using database knowledge.

## 6.6 Justification of the Energy Function

We design and perform an empirical statistical test to show that the true signaling pathway structure has the lowest energy in the feasible set. Given the true signaling pathway structure $(\overline{X}, \overline{\Theta})$, we randomly select $N$ feasible structures and compute the empirical $P$-value $M/N$, where $M$ is the number of structures with energy lower than that of $(\overline{X}, \overline{\Theta})$. The true signaling pathway structure has the lowest energy if the empirical $P$-value is zero. We also perform the above test for a randomly selected feasible structure and expect the empirical $P$-value to vary in the interval [0 1].

## 6.7 Gene Set Simulated Annealing (GSSA)

For the search procedure, we define the neighborhood of a signaling pathway structure $(\overline{X}, \overline{\Theta})$ as the set of $\sum_{i=1}^{m}(L_i-2)!$ structures obtained by randomly permuting the orders of $L_i-2$ intermediate genes in the $i^{th}$ IF $(X_i, \Theta_i)$, keeping the remaining $m-1$ IFs in $(\overline{X}, \overline{\Theta})$ fixed, for each $i = 1, \ldots, m$. This definition justifies the term 'neighbor' as only one IF in the given structure is perturbed at a time. Moreover, if we start our search from a feasible structure, the algorithm is guaranteed to take jumps within the feasible set of candidate structures having the same degree distribution as the true signaling pathway. The above definition also satisfies all the properties of a neighborhood presented in [44]. We choose the standard cooling schedule, which at the $k^{th}$ stage is defined as

$$T_k = \frac{c}{\log(k+1)},\ k = 1, 2, \ldots, \tag{6.5}$$

where $c > 0$ is constant and is referred to as *cooling schedule constant*. The choice of $c$ is often problem specific. Indeed, a small value of $c$ may lead SA to get trapped in a local solution, whereas a large value may slow down its speed of convergence. The above cooling schedule has been used to study the convergence properties of a general simulated annealing approach [47]. The probability with which the algorithm accepts a move from a current structure $(\overline{X}, \overline{\Phi})$ to a neighboring structure $(\overline{X}, \overline{\Psi})$ is called the *acceptance probability* [25] and is defined as

$$\min\{1, \exp(\mathcal{E}(\overline{X}, \overline{\Phi}) - \mathcal{E}(\overline{X}, \overline{\Psi})/T)\} \tag{6.6}$$

where $T$ represents the current temperature value, which at the $k^{th}$ iteration is given by Eq. 6.5. Note that the algorithm may accept to move to a worse point in order to avoid getting trapped in a local solution. In Algorithm 6.1, we present the pseudo-code of GSSA. Algorithm 6.1 takes an IFGS compendium as input and returns a list of IFs, which are combined to represent the optimal signaling pathway structure.

## 6.8 Description of the Case Studies

### 6.8.1 Case Study I: Using Signaling Pathway Structures in KEGG

#### Data

We first performed a proof-of-principle study to validate the performance of GSSA in inferring the true signaling mechanisms. In this study, we considered a compendium of gene sets sampled from the true signaling pathway structure. We developed a path sampling algorithm (see Appendix A.8) for obtaining a collection of true IFs from a given signaling pathway structure. The IFGSs were simulated by randomly permuting the locations of intermediate genes within each IF by keeping the pair of terminal nodes fixed. We applied this algorithm individually on 120 non-metabolic pathways in the KEGG database [67, 68] resulting in 120 IFGS compendiums. From each compendium, we pruned IFGSs of lengths 2 and 3. Such

IFGSs represented true edges and true IFs, respectively. Further, we only used compendiums which contained at least 5 IFGSs. Using this procedure, we obtained 83 non-empty IFGS compendiums comprising of IFGSs sampled from KEGG. Since each compendium was derived from a specific KEGG pathway structure, IFGSs in a given compendium shared the same pathway membership. In the derived compendiums, the number and lengths of IFGSs varied in the ranges of $5-723$ and $4-13$, respectively. These compendiums served as input to GSSA.

**Comparative Analysis**

We considered two Bayesian network approaches: K2 [26] and Metropolis-Hastings or MH [100] implemented in BNT [101]. We used both BIC and Bayesian scoring functions to infer Bayesian networks. In the case of K2, the maximum number of parents allowed for a node was set at 3. In each run of MH, the first 1000 samples were collected for a manageable computational complexity, and the structure giving the highest F-Score was selected for comparison. Results were averaged from 10 independent runs of each algorithm.

We also compared GSSA with MI based approaches. Binary discrete data corresponding to an IFGS compendium served as input to MI based algorithms. In the comparison, we treated the true signaling pathway structure and the one inferred by GSSA as undirected. To infer MI networks, we used ARACNE [88], C3NET [8], CLR [35], MRNET [96] and RN [19] approaches implemented in the C3NET [8] and MINET [97] packages available from R/Bioconductor. We used the empirical MI estimator to estimate mutual information matrix. We did not observe a significant difference by employing other estimators defined for discrete random variables. The final normalized network was compared with the true structure at several threshold values. For each algorithm, the structure maximizing the F-score was considered as the inferred structure. We observed an overall better performance of ARACNE, when the DPI threshold parameter $\epsilon$ was set at 0 [97]. Therefore, we set $\epsilon = 0$ in Case Study I.

### 6.8.2 Case Study II: Using *E. coli* Data Sets

We also compared the performance of various approaches using 4 benchmark *E. coli* data sets, considered in Section 5.6.2. These data sets are available from DREAM3 network challenges in the DREAM initiative [86, 87, 117]. Since we do not know the parameters that maximize the performance of a network inference algorithm in real-world studies, we used a standard procedure to infer the underlying network from each of the 4 data sets.

To infer MI networks, we applied copula transform on each data set for a stable estimation of mutual information [8]. We did not observe a significant difference in the performance of MI based algorithms without using copula transform. We used the empirical Gaussian estimator implemented in the C3NET package [8, 9] to estimate the mutual information matrix (MIM). MIM was given as input to CLR and MRNET. We used a MI threshold on the resulting matrix to remove non-significant edges. We set the MI threshold as the average of values in the upper triangular part of the inferred matrix. In the case of ARACNE and C3NET, we first used a MI threshold on MIM. The resulting matrix served as input to both ARACNE and C3NET. MI threshold was set as the average of values in the upper triangular part of MIM, which is also the default cut-off used in C3NET package. The DPI threshold parameter for ARACNE was set at 0.1 [8].

In the case of GSSA, we considered 4 IFGS compendiums derived in Section 5.6.2, which comprised of 47, 45, 45 and 49 IFGSs, respectively. For each compendium, we used GSSA to explore the search spaces formed by considering all possible gene orderings of gene sets present in it.

In the case of Bayesian network methods, we applied K2 and MH on both continuous and discretized *E. coli* data sets at different settings of parameters.

#### Comparative Analysis

We tested the performance of GSSA, Bayesian network and MI based methods using the above parameter settings. Performances were measured in terms of F-score (Section 5.6.1).

### 6.8.3 Case Study III: Pathways Reconstruction in Breast Cancer Cells

In this study, we showcase two context-specific signaling pathways, ERBB and PMOM (Progesterone-mediated oocyte maturation), activated in breast cancer. We considered 87 genes participating in the ERBB signaling pathway and 35 genes in the giant connected component (GCC) of the PMOM pathway from the KEGG database. We analyzed 299 clinical breast cancer tissue gene expression profiles from the Affymetrix HG-U133 plus 2.0 platform considered in Section 5.6.3. This resulted in two data sets of size $87 \times 299$ and $35 \times 299$ corresponding to the genes in the two pathways. To derive IFGS compendiums, we discretized each data set using binary labels.

Specifically, we derived two IFGS compendiums, Compendiums I and II, corresponding to the genes in the ERBB and PMOM pathways, respectively, with a minimum of 4 component genes in each IFGS. As the majority of IFGSs ($\sim 90\%$ in Compendium I and $\sim 94\%$ in Compendium II) were comprised of $4-9$ genes, such samples provided a good compromise between the overlapping among IFGSs and the time for convergence. This resulted in two compendiums with 204 and 96 IFGSs, respectively. We assigned the end nodes for each context-specific IFGS using the hierarchial representation of genes in different layers of the generic ERBB and PMOM pathway structures in the KEGG database. The hierarchial representation of a signaling pathway can be visualized using Cytoscape [131]. Within each IFGS, a gene lying in the upper most and a gene in the lower most layer were considered as the two end nodes. It is worth mentioning here that layering information accounts for the gene orderings at a very crude level because (1) The derived IFGSs do not necessarily correspond to signaling events already reported in KEGG (2) No prior knowledge of edges in the two KEGG structures was used. Lists of genes in the two compendiums along with their hierarchial arrangements in the different layers of the two KEGG pathways have been presented in Table 6.1.

| | Genes |
|---|---|
| Layer1 | EGF, TGFA, AREG, EREG, BTC, HBGEF, ERBB2, NRG1, NRG2, NRG3, NRG4 |
| Layer2 | EGFR, ERBB3, ERBB4 |
| Layer3 | SRC, CBL, CBLC, CBLB, NCK1, NCK2 PLCG1, PLCG2, STAT5A, STAT5B, SHC1, SHC2, SHC3, SHC4, CRK, CRKL |
| Layer4 | PTK2, PAK1, PAK2, PAK3, PAK4, PAK6, PAK7 CAMK2A, CAMK2B, CAMK2D, CAMK2G, PRKCB PRKCA, PRKCG, GRB2, ABL1,ABL2 |
| Layer5 | MAP2K4, MAP2K7, SOS1, SOS2, GAB1 |
| Layer6 | MAPK8, MAPK9, MAPK10, NRAS, HRAS, KRAS, PIK3R2 PIK3CA, PIK3R3, PIK3R5, PIK3CB, PIK3CD, PIK3R1, PIK3CG |
| Layer7 | JUN, ARAF, BRAF, RAF1, AKT1, AKT2, AKT3 |
| Layer8 | MAP2K1, MAP2K2, BAD, MTOR, CDKN1A, CDKN1B, GSK3B |
| Layer9 | MAPK1, MAPK3, EIF4EBP1, RPS6KB1, RPS6KB2 |
| Layer10 | ELK1, MYC |

| | Genes |
|---|---|
| Layer1 | HSP90AA1, HSP90AB1, PLK1, SPDYA, SPDYC |
| Layer2 | MOS, CDK2, CDC25A, CDC25B, CDC25C |
| Layer3 | MAP2K1 |
| Layer4 | MAPK1, MAPK3 |
| Layer5 | RPS6KA1, RPS6KA2, RPS6KA3, RPS6KA6 |
| Layer6 | BUB1, PKMYT1 |
| Layer7 | MAD1L1, MAD2L1, MAD2L2 |
| Layer8 | FZR1 |
| Layer9 | ANAPC1, ANAPC2, ANAPC4, ANAPC5, ANAPC7, ANAPC10, ANAPC11, ANAPC13, CDC16, CDC23, CDC26, CDC27 |

Table 6.1: The hierarchial arrangement of 87 genes from the ERBB signaling pathway (Upper Panel) and 35 genes from the PMOM pathway (Lower Panel) available from the KEGG database [67, 68]. These representations can be visualized using Cytoscape [131].

## 6.9 Performance Evaluation

### 6.9.1 Using IFGSs Derived from Signaling Pathway Structures in KEGG

We began by examining that the true signaling pathway structure has the lowest energy in the feasible set. We considered two collections of feasible structures. The first collection

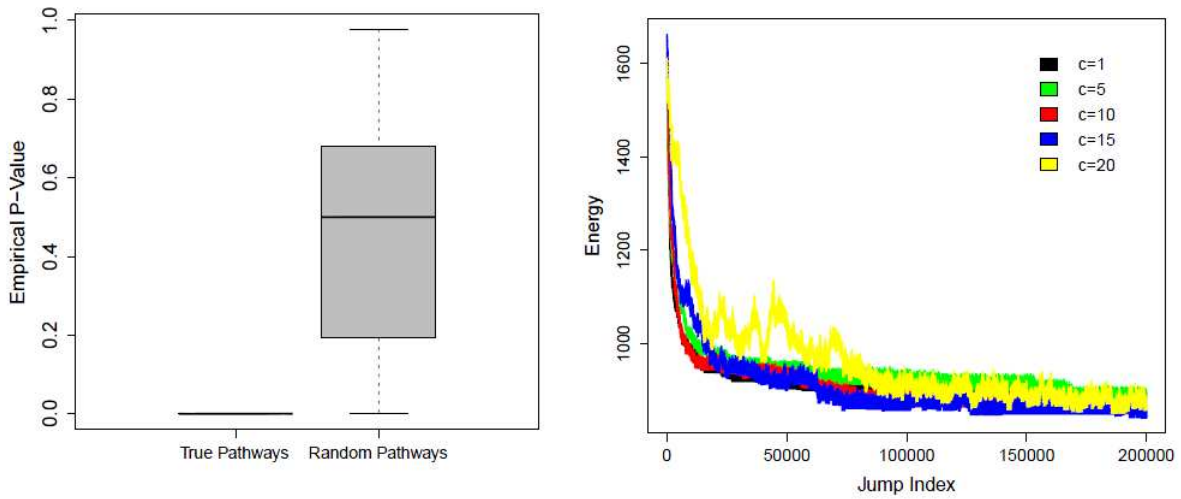


Figure 6.1: Left Panel: Empirical P-Values computed for true signaling pathway structures (Left) and randomly selected feasible pathway structures (Right) corresponding to 83 IFGS compendiums derived from the KEGG pathways; Right Panel: Energy values computed by varying the initial structure and cooling schedule constants for a total of $2 \times 10^5$ jumps. The IFGS compendium was derived from the generic vascular smooth muscle contraction pathway in KEGG.

comprised of all 83 signaling pathway structures constructed from the true IFs. The second collection contained 83 randomly selected structures, one from each of the 83 feasible sets. The left panel of Fig. 6.1 presents the empirical $P$-values calculated for each structure in the two collections, where we fixed $N = 1000$ (see Methods). We observed that the empirical $P$-value for each of the 83 true structures was always zero while it fluctuated in the interval (0 1) in the case of randomly selected feasible structures. This justified the choice of the energy function used in our algorithm.

Since the computational complexity of GSSA was quite manageable for the derived compendiums, we fixed the number of jumps in a single run of GSSA at $2 \times 10^5$. Based on our experiments, we chose to fix $c = 10$ throughout to compromise between the problem of getting stuck in a local solution and the time needed for convergence. Fig. 6.1 (Right Panel) presents the energy values from five independent runs of GSSA with different cooling schedule constants and different initial structures randomly chosen from the feasible set. It can be observed from Fig. 6.1 that at a later stage, the energy values obtained using $c = 10$

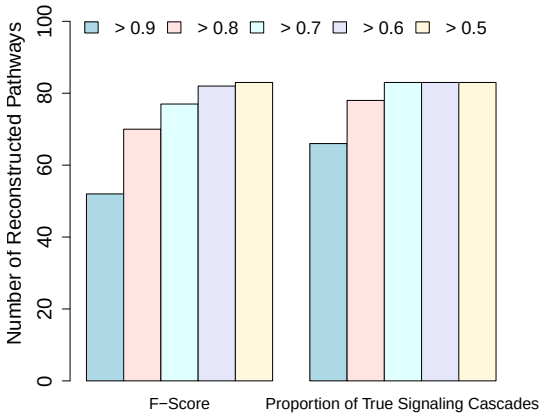


Figure 6.2: The performance of GSSA in reconstructing true signaling cascades and signaling pathway structures corresponding to 83 IFGS compendiums derived from the KEGG database.

are very close to the ones calculated from other settings.

We summarize the performance of GSSA in terms of F-Score averaged over 10 independent runs. Fig. 6.2 demonstrates the performance of GSSA in reconstructing the true signaling mechanisms using each of the 83 IFGS compendiums. On the left side of Fig. 6.2, we have plotted the number of structures among 83 reconstructed structures, with a certain minimum F-Score. On the right, we consider the proportion of signaling cascades accurately inferred by our algorithm in each compendium. The feasibility and validity of GSSA is evident from the high F-Scores and the high proportion of accurately inferred signaling cascades.

In Fig. 6.3, we present the results from a comparative study performed using each of the 83 IFGS compendiums. We observe a significantly better performance of GSSA in recovering the true structure compared with the Bayesian network and MI based approaches. Fig. 6.3 demonstrates the strength of GSSA in inferring signal cascading mechanisms.

In Fig. 6.4, we present a signaling pathway structure inferred by our approach. Structures on the upper and lower panels correspond to the true and inferred signaling pathway structures, respectively. The black (solid) and blue (dashed) edges represent true

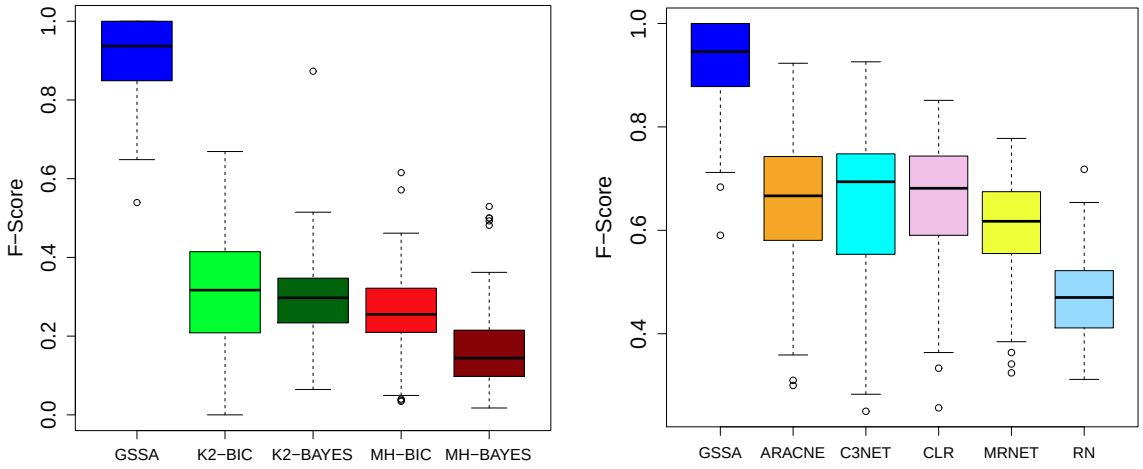


Figure 6.3: Comparison of GSSA with the Bayesian network approaches K2 and MH using BIC and Bayesian score functions (Left Panel) and with MI based approaches (Right Panel).

positives and false positives, respectively. Fig. 6.4 demonstrates high precision and recall in the structure reconstructed by GSSA, resulting in a high F-Score.

### 6.9.2 Using IFGSs Derived from the *E. coli* Data Sets

Since we observed very low sensitivity values using Bayesian network methods, we present the result from comparison between GSSA and MI based methods. In Fig. 6.5, we plot the performance of GSSA and MI based network inference methods in terms of the precision ratio, which is the ratio of the precision from GSSA and the one from MI based methods. A precision ratio more than 1 indicates a better precision by GSSA. For each *E. coli* data set, we observed a higher precision from GSSA, compared with MI based network inference methods.

### 6.9.3 Using IFGSs Related to the ERBB and PMOM Signaling Pathways

We inferred two breast cancer specific signaling pathway structures using the derived compendiums. To evaluate the performance of GSSA, we first utilized the structures of ERBB

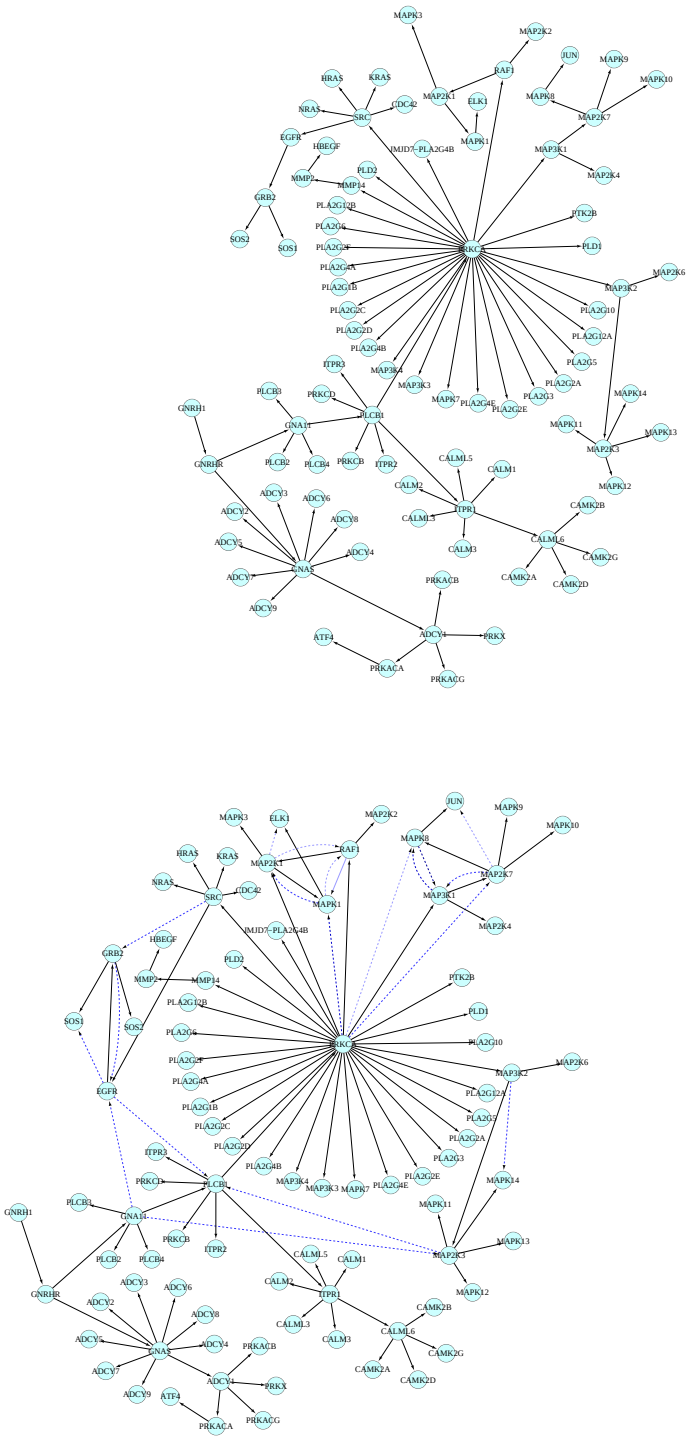


Figure 6.4: An example showcasing the performance of GSSA in recovering the true structure using the IFGS compendium derived from the GnRH signaling pathway in KEGG database. Structures on the upper and lower panels represent true and inferred signaling pathways, respectively. The black (Solid) and blue (Dashed) edges represent true positives and false positives, respectively. Figures were generated using Cytoscape [131].

and PMOM signaling pathways in the KEGG database [67,68]. Considering that the direction of an information flow is often from an upper layer to a lower one in the hierarchial representation of a signaling pathway, and the real-world gene sets correspond to partially observed signaling events, at the minimum we expected a larger number of inferred edges between genes in upper layers to genes in lower layers in the hierarchial representation of the two KEGG pathway structures. Indeed, we verified that nearly 76% and 89% of the inferred edges follow this hierarchy, i.e. no parent came from a layer lower than the one for its child. This observation indicates that for a vast majority of inferred signaling mechanisms, the flow of information was from an upper layer to a lower one.

In the upper panel of Fig. 6.6, we present a few reconstructed signal transduction events which correspond to complete or partial linear signal cascades already reported in the ERBB and PMOM pathway structures in the KEGG database. In the lower panel of Fig. 6.6, we present a partial view of the two reconstructed signaling pathways with solid edges representing complete or partial linear signal cascades already reported in the ERBB and PMOM signaling pathways in the KEGG database, whereas dashed edges follow the hierarchy of these structures and can be viewed as predictions. While the figures do not attempt to portray a comprehensive view of signaling pathways, GSSA algorithm has the potential to uncover biologically relevant mechanisms that have not been previously considered or understood.

ERBB/HER family receptors play important roles in many types of cancer including breast cancer. Dysregulation/mutation in the epidermal growth factor receptor (EGFR) and ERBB2 (HER2) have been known to promote angiobenesis and metastasis in breast cancer [83,104]. Some known signaling cascades that contribute to breast cancer progression include RAF/MEK/ERK and PI3K/PDK1/AKT signaling pathways that regulate apoptosis and cell cycle. These signaling events are reflected in the edges depicted in the upper left panel of Fig. 6.6. For instance, in breast cancer ERBB2/HER2 receptor can constitutively activate the PI3K/PDK1/AKT cascade and the downstream effector, the mammalian target

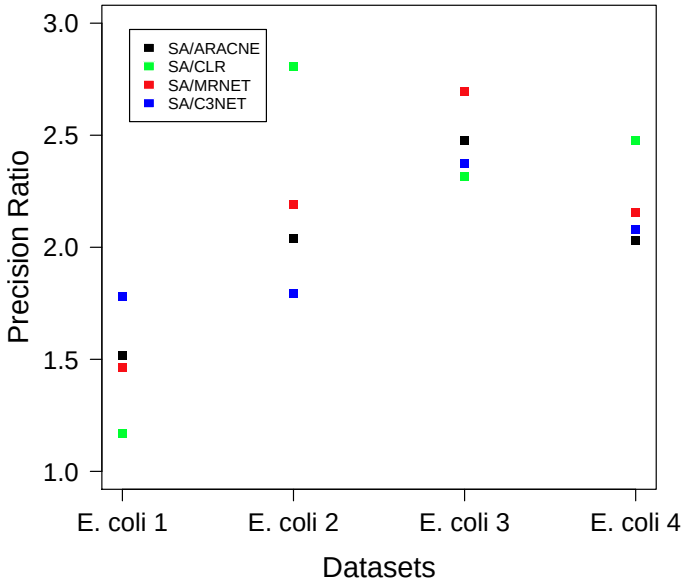


Figure 6.5: Comparison of GSSA and the MI based methods in terms of precision ratio, which is the ratio of the precision from GSSA and the one from MI based methods. We used 4 *E. coli* benchmark data sets available from the DREAM initiative.

of rapamycin (MTOR). This known signaling cascade is conformed as a direct action between ERBB2/HER2 and MTOR in the lower left panel of Fig. 6.6.

In the lower left panel of Fig. 6.6, the reconstructed ERBB signaling pathway revealed a previously unknown direct link from ERBB3 to ARAF. ARAF (A-Raf proto-oncogene serine/threonine-protein kinase) is known to phosphorylate and activate MEK1 (MAP2K1) and MEK2 (MAP2K2), leading to suppression of apoptosis in cancer cells [123]. However, the possible role of ERBB3 as its upstream regulator is a novel implication that clearly warrants further investigation. In addition, PI3K family members are known to be the downstream targets of EGFR and ERBB2/HER2, but not ERBB3 [22]. Thus, the direct link between ERBB2 and PI3K inferred by GSSA is in accordance with the previously established results. The direct link between ERBB3 and PIK3R3, on the other hand, suggests a potential role of ERBB3 receptor tyrosine kinase in breast cancer. A major clinical challenge of breast cancer treatment is acquired resistance to hormone therapy as the tumor develops

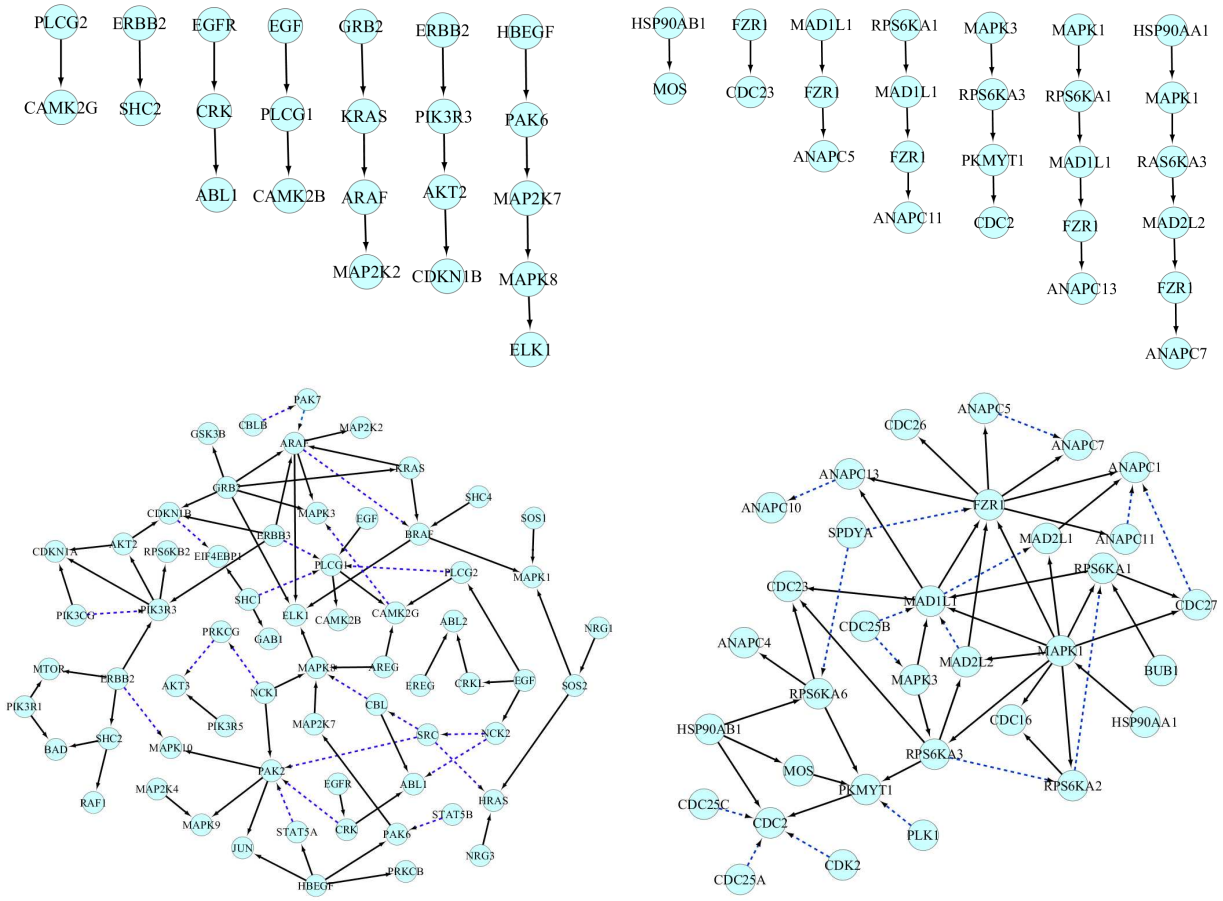


Figure 6.6: Upper Panel: Linear cascading events inferred by GSSA which correspond to complete or partial linear signaling events already reported in the ERBB (Upper Left) and PMOM (Upper Right) pathways in KEGG; Lower Panel: Partial view of the breast cancer signaling pathways, ERBB (Lower Left) and PMOM (Lower Right), inferred by GSSA. A solid edge represents that a complete or partial linear signaling event between parent and child node has been recognized in the ERBB and PMOM structures in the KEGG database, whereas dashed edges follow the hierarchial arrangements of these structures.

alternative survival signaling such as enhanced cross-talk between the estrogen receptor (ER) and ERBB1/ERBB2 [129]. Thus combinatorial therapeutic intervention targeting both ER and ERBB2 (HER2) is currently under intensive clinical studies [79, 80, 110]. Revelation of the novel link between ERBB3 and PI3K family proteins is significant because it represents yet another adaptive pathway in breast cancer that needs to be fully understood in order to develop more effective regimen blocking this survival signaling.

In the case of PMOM pathway (lower right panel of Fig. 6.6), we show a highlighted role of the Fizzy protein (FZR1/CDC20) in breast cancer. It is an indication that the ubiquitin ligase activity of the anaphase promoting complex (APC) plays an important role in breast cancer progression. Previous studies have established an association between APC and FZR1 [141] implicating FZR1 regulation of ANAPC isoforms 1, 2, 4, 5, 7, and 10. We observe additional regulation mechanisms involving ANAPC 11 and 13, apparently in a way specific to breast tumor tissues. The reconstructed PMOM signaling pathway also reveals a novel direct action of mitogen-activated protein kinase 1 (MAPK1) upon FZR1. The MAP kinase cascade is associated with the control of cell cycle progression, but in a manner that is far upstream of FZR1-mediated APC. It is possible that this direct action may be a result of the non-genomic signaling of progesterone [14] that rapidly and constitutively activates the MAP kinase signaling cascade in breast cancers that are estrogen receptor (ER) positive but progesterone receptor (PGR) negative.

If experimentally validated and mechanistically elucidated, the novel activation of FZR1 by MAPK1 will have important outcomes in breast cancer research. For example, studies can be designed to investigate if inhibiting the kinase can block FZR1-mediated APC, and if any effector proteins are involved in this signaling cascade. Such studies can be driven by hypotheses generated from GSSA-based reconstruction of signaling pathways, and can lead to the discovery of new biomarkers as potential diagnostic, prognostic, or therapeutic targets for breast cancer.

## 6.10 An Alternative Approach: Gene Set Genetic Algorithm (GSGA)

The discrete optimization problem considered in Eq. 6.1 can also be addressed by utilizing the frameworks of other popular search techniques, such as genetic algorithm (GA) [28,53,98]. We performed a preliminary study for the structural inference of signaling pathways under the settings of GA, which we describe below.

GA is a population based search strategy, which starts from an *initial population* of points (signaling pathway structures) from the feasible set. Points in the feasible set are encoded as strings of symbols of equal lengths and are called *chromosomes*. GA proceeds iteratively, where a new population is created from the current population using the operations referred to as *cross-over* and *mutation*. At each iteration, GA aims to create a population with average objective function value, which is higher than the one for the previous population. The objective function value of a chromosome is called its *fitness*. Various steps in the proposed gene set based genetic algorithm, GSGA, are as follows:

Problem Formulation We formulate the discrete optimization problem in Eq. 6.1 as a maximization problem

$$\max_{(\overline{X},\overline{\Theta})\in\mathcal{F}_{\overline{X}}} f(\overline{X},\overline{\Theta}) \tag{6.7}$$

where $f$ represents the fitness of a signaling pathway structure $(\overline{X},\overline{\Theta})$ and is defined as

$$f(\overline{X},\overline{\Theta}) = \sum_{i=1}^{m} \log \ell(X_i,\Theta_i). \tag{6.8}$$

The Representation Scheme We encode each signaling pathway structure in $\mathcal{F}_{\overline{X}}$ as a chromosome. To do this, we first enumerate the orderings associated with each of the IFGSs $X_i,\ i=1,\ldots,m$ individually and label the corresponding IFs based on the enumeration. We then concatenate the labels of the IFs which define the given signaling pathway structure. For instance, if a signaling pathway structure is defined in terms of three IFGSs $X_1$, $X_2$ and $X_3$, with lengths 5, 7 and 6, respectively, and the ordering index 4, 3 and 5 associated

**Algorithm 6.2** Gene Set Genetic Algorithm

**Input:** IFGSs $X_i,\ i = 1, \ldots, m$, population size $s$, cross-over probability $p_c$, mutation probability $p_m$, elitism proportion $p_e$, number of generations $J$.
**Output:** The reconstructed signaling pathway structure.
**Initialization:** At $k = 0$, randomly select a population $P^{(0)}$ of size $s$ from $\mathcal{F}_{\overline{X}}$. If $(\overline{X}, \overline{\Theta}^{(0)})$ is the structure with the maximum fitness in $P^{(0)}$, let BestNetwork $= (\overline{X}, \overline{\Theta}^{(0)})$ and BestFit $= f(\overline{X}, \overline{\Theta}^{(0)})$.
**for** $k = 1, \ldots, J$ **do**
  Let $P^{(k)} = \{\}$.
  **if** $p_e > 0$ **then**
    Put a total of $n_e$ chromosomes from $P^{(k-1)}$ with the first $n_e$ highest fitness values into $P^{(k)}$, where $n_e = \lfloor p_e * s \rfloor$. Let $C^{(k-1)}$ be the set of the remaining chromosomes in $P^{(k-1)}$.
  **else**
    $C^{(k-1)} = P^{(k-1)}$.
  **end if**
  Form a mating pool $M^{(k-1)}$ from $C^{(k-1)}$ using a tournament scheme.
  Apply cross-over on the chromosomes in $M^{(k-1)}$ with probability $p_c$. Update $M^{(k-1)}$.
  Apply mutation on the chromosomes in $M^{(k-1)}$ with probability $p_m$. Update $M^{(k-1)}$.
  Include the chromosomes of $M^{(k-1)}$ into $P^{(k)}$.
  **if** $(\overline{X}, \overline{\Theta}^{(k)})$ is the structure with the maximum fitness in $P^{(k)}$ and BestFit$< f(\overline{X}, \overline{\Theta}^{(k)})$ **then**
    BestNetwork $= (\overline{X}, \overline{\Theta}^{(k)})$.
    BestFit $= f(\overline{X}, \overline{\Theta}^{(k)})$.
  **end if**
**end for**
Return BestNetwork and BestFit.

with the IFGSs, respectively, define the signaling pathway structure, then the chromosome is recognized by the symbol 435. Each structure in $\mathcal{F}_{\overline{X}}$ can be encoded as a chromosome in a similar way.

Mating Pool From a given population $P^{(k)}$, which is a set of signaling pathway structures, we generate a mating pool $M^{(k)}$ using a *tournament scheme*. To do this, we randomly select two chromosomes and put the chromosome with a better fitness value into the pool. If the size of the population is $s$, we repeat the tournament $s$ times.

Cross-over In cross-over, we select a pair of parent chromosomes from the mating pool and exchange a pre-specified number of IFs between them. A given proportion of the chromosomes in $M^{(k)}$ go through cross-over.

Mutation In mutation, we take each chromosome from $M^{(k)}$ and randomly permute the

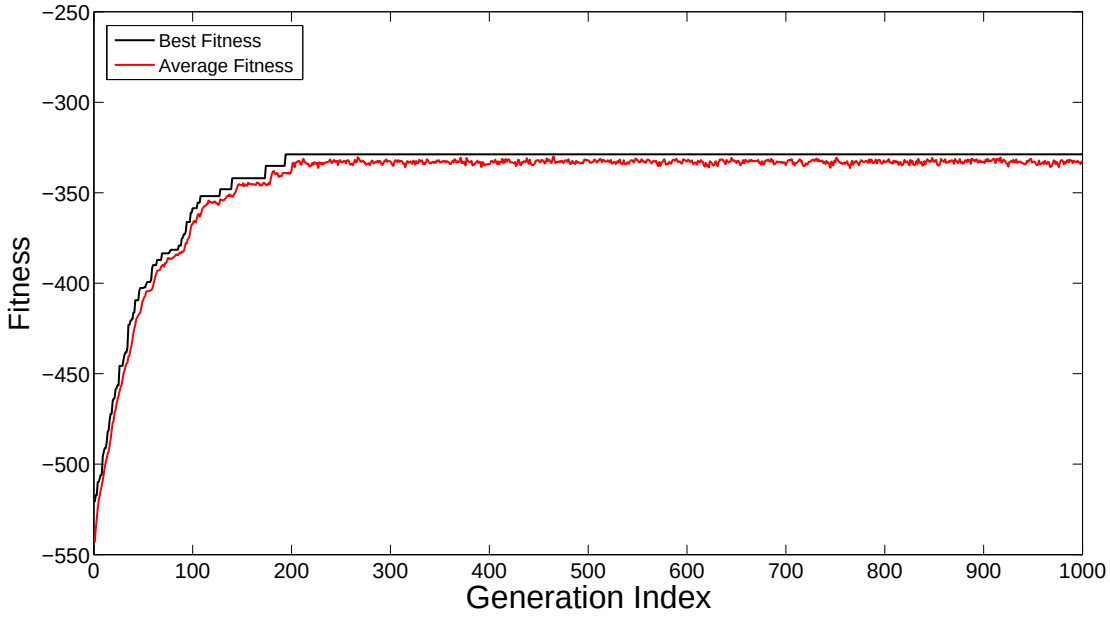


Figure 6.7: Convergence of GSGA to the global solution using the IFGS compendium derived from the *E. coli* network considered in Section 5.6.1.

ordering of each of the $m$ IFs with a very small probability.

Elitism The mating pool $M^{(k)}$ obtained after applying cross-over and mutation represents the new population or *generation* $P^{(k+1)}$. However, we can further restrict a pre-specified proportion of the chromosomes, say a total of $n_e$ in number, with the first $n_e$ highest fitness values in the current population to transfer to the next population, without going through cross-over or mutation. This scheme is referred to as elitism.

GSGA iteratively repeats the above process starting from $P^{(k+1)}$, until a specified number of generation is reached. GSGA has been presented in Algorithm 6.2.

In Fig. 6.7, we show the convergence of Algorithm 6.2 to the global solution, where we used the IFGS compendium derived from the *E. coli* network considered in Section 5.6.1. We set $s = 50$, $p_e = 0.25$, $p_m = 0.01$, $p_c = 0.25$ and $J = 1000$.

## 6.11 Discussion

In this chapter, we presented a simulated annealing algorithm to infer the optimal signaling pathway structures from gene sets related to the pathways. By hypothesizing the underlying signaling pathway structure as an ensemble of overlapping signaling cascades, we formulated its reconstruction from gene sets corresponding to signaling cascades as a discrete optimization problem. Throughout we treated gene sets as random samples from a first order Markov chain model and their orders as random. We also presented a new energy function to measure the optimality of a signaling pathway structure.

In Case Study I, performance evaluation using 83 gene set compendiums derived from KEGG pathways demonstrated that GSSA could recover the underlying structures more efficiently than other contemporary approaches. In Case Study II, reconstruction of benchmark *E. coli* networks and in Case Study III, breast cancer specific reconstruction of two signaling pathway structures from the KEGG database further proved the advantages of using GSSA in real-world scenarios.

The worst-case running time of GSSA is $\mathrm{O}(JmL)$, where $J$ is the number of jumps, $m$ is the number of IFGSs and $L$ is the maximum length of an IFGS in the given compendium. We refer to Appendix A.11 for a detailed discussion on the computational complexity of GSSA. Overall, GSSA benefits from a manageable computational load compared with search heuristics such as sampling based Meteropolis-Hastings algorithm used in the inference of Bayesian networks. MI based algorithms are computationally more efficient that GSSA and Bayesian network methods, however, they are suitable for inferring undirected pairwise dependencies.

Gene set based reconstruction of signaling pathway structures offers a simple and flexible approach with numerous possibilities of extension. For instance, we showed that the problem discussed in this chapter can also be addressed by utilizing the framework of genetic algorithm.

# Chapter 7

# Conclusion and Future Works

In this dissertation, we presented a series of multivariate approaches for inferring gene interaction and regulation patterns from molecular profiling data. The overall work was composed of two parts. In the first part, we presented models and algorithms leading to a reliable discovery of gene clusters or pathway components. Our approach was to learn an optimal correlation structure from replicated complete and incomplete molecular profiling data. In the second part, we considered the problem of inferring signal transduction mechanisms in a given pathway component. We addressed the problem by treating gene sets corresponding to signal transduction activities as the basic building blocks of the underlying signaling pathway structure. We comprehensively examined the performance of our approaches using simulated and real-world data sets.

In particular, the presented research makes original contributions by addressing the following challenges in systems biology:

*Correlation-based pattern discovery from replicated molecular profiling data*: Outcome of any bioinformatics analysis is directly influenced by the quality of experimental data. It is well-known that molecular profiling measurements produced by high throughput data acquisition platforms are often contaminated with excessive noise. Replication is frequently used in such cases to account for the noise and to achieve a reliable discovery of the underlying biomolecular activities. However, the analysis of replicated molecular profiling measurements is challenging for the following reasons: (1) there often exists a large variation in the magnitudes of replicated measurements (2) the replication mechanism used in underlying experimental design may be known *a priori*, however, a data analysis method may fail to accommodate

this information (3) in several cases, replicated measurements contain a small to large percentage of missing values. Therefore, it is necessary to develop effective methodologies for gaining insights from both replicated complete and replicated incomplete molecular profiling data.

The existing approaches to correlation estimation do not automatically accommodate replicated measurements. Often, an *ad hoc* step of data preprocessing by averaging (either weighted, unweighted or something in between) is used to reduce the multivariate structure of replicated data into a bivariate one [59,151,152]. Averaging may create a strong bias while reducing the variance among replicates of a gene. Averaging may also wipe out important patterns of small magnitudes or cancel out opposite patterns of similar magnitudes, resulting in a significant amount of information loss. Multivariate approaches must be designed to sufficiently exploit each replicated measurement individually. This is the main motivation behind Chapters 2-4 presented in this dissertation. Throughout these chapters, we treated each replicate as a variable by assuming that data were independently and identically distributed samples from multivariate normal distribution(s). Specifically,

- In Chapter 2, we introduced an informed-case model [4, 161] for estimating the correlation structure of a gene set with replicated and complete molecular profiling data. Informed-case model generalizes previously known blind-case or parsimonious model [1, 4, 158] by accommodating prior knowledge of replication mechanisms. Indeed, the number of biological and technical replicates used in underlying experimental design are known in many cases. While the blind-case model imposes the same correlation parameter for different biological replicates of a gene, informed-case model allows them to be different.

- In Chapter 3, we further generalized informed-case model by designing a two-component mixture model [1, 4]. The underlying idea was to shrink the correlation structure of a gene set with replicated and complete measurements between a constrained correlation structure and an unconstrained one. The constrained correlation structure was the one

given by blind-case model, whereas the unconstrained correlation structure was free from any parameter constraints.

- For the estimation of correlation structure from replicated and incomplete molecular profiling data, we developed an Expectation-Maximization (EM) algorithm in Chapter 4 [161]. EM algorithm iterates between the E step and the M step until convergence. The E step computes the expected values of the sufficient statistics for underlying multivariate normal distribution given by either blind-case or informed-case model, whereas the M step updates the current estimates of the model parameters.

By utilizing correlation distance as metric, we used the above multivariate models and algorithms for clustering real-world replicated data sets with both complete and incomplete measurements. Gene clusters are often interpreted as pathway components, which comprise of a group of molecules (usually proteins) upstream of transcription factors. Activation of a pathway initiates sequences of signal transduction which affect gene expressions via downstream transcription factors. Inference of directed network topology representing signal transduction activities in a pathway is a major challenge in systems biology. We developed two novel algorithms to address this challenge.

*Reconstruction of signaling pathway structures*: We dealt with the problem of reconstructing signaling pathway structures by utilizing a compendium of gene sets related to the pathway. Indeed, the advent of systems biology has been accompanied by the blooming of network reconstruction algorithms, many of which treat gene pairs as the basic building block of the signaling pathways and reconstruct the underlying structure by simultaneously detecting co-expressed gene pairs using molecular profiling data [19, 35, 88]. This type of approaches enjoy simplicity and a much alleviated computational load but gene pairs do not represent the entire signal transduction events. Other approaches heuristically search for the higher scored network structure(s), such as bayesian networks [26, 37, 130]. Many network structures may be found to be statistically plausible, but similar to the gene pairs they do not necessarily represent the real signal transduction mechanisms. Moreover, the computational

load of searching for a higher scored network is prohibitively high [24, 122] and a number of assumptions on the network structures have to be made.

We hypothesized a signaling pathway structure as an ensemble of several overlapping signal transduction events with a linear arrangement of genes in each event. Gene sets, in our context, referred to sets of genes participating in directed chains of signal transduction. We proposed to infer the true signaling pathway structure by inferring the order of genes in each gene set and combining the inferred chains of signal transduction into a single unit. Throughout, we treated unordered gene sets as random samples from a first order Markov chain model and their orders as random. Our motivation of considering a gene set based approach for the structural inference of signaling pathways falls into many categories. For instance, a gene set based approach can more naturally incorporate higher order signaling mechanisms as opposed to pairwise interactions. In comparison to continuous molecular profiling data, gene sets are more robust to noise and facilitate data integration from multiple data acquisition platforms. We proposed two novel gene set based algorithms to achieve our goal. Particularly,

- In Chapter 5, we translated our goal of signaling pathway structure inference into drawing samples of signaling pathway structures sequentially from the joint distribution of gene sets followed by summarizing the most likely structure from the sampled structures. We developed a stochastic algorithm, Gene Set Gibbs Sampler (GSGS) [2], under the Gibbs sampling framework [40, 41] to achieve our goal. In the GSGS framework, we sample a signaling pathway structure by sampling an order for each gene set from a conditional distribution defined by the remaining gene sets in the compendium.

- In Chapter 6, we provided a search strategy, as opposed to sampling strategy used in GSGS, for learning the optimal signaling pathway structure from gene sets [3]. We first formulated the structural inference of signaling pathways from gene sets into a discrete optimization problem and then presented a simulated annealing algorithm [74], GSSA, to infer the ordering of genes in the gene sets. GSSA mimics the physical process

of heating and then cooling down a substance slowly to obtain a strong crystalline structure, by annealing gene sets to infer signaling cascades characterizing the optimal signaling pathway structure.

The past decade has witnessed a significant progress in the computational inference of biological networks. A variety of approaches in the form of network models and algorithms have been proposed to understand the structure of biological networks at both global and local levels. While the grand challenge in a global approach is to provide an integrated view of the underlying biomolecular interaction and regulation mechanisms, a local approach focuses on the study of fundamental domains representing functional units or biological pathways. However, the existing computational approaches often rely on unrealistic biological assumptions and do not sufficiently exploit the potential of molecular profiling data available in diverse forms. Gene set based approaches discussed in this dissertation offer a fresh perspective to explore the structural organization of biological networks with several possibilities of extensions. In particular, our current study can be further extended in the following directions:

*Discovery of pathway components:* A reliable discovery of pathway components is the first major step towards understanding signal transduction mechanisms. This step relies on the strength of a computational approach to fully exploit the complex dependency structure underlying molecular profiling data. In this dissertation, we followed the path of correlation-based patten discovery, which serves as a bridge between replicated complete and incomplete molecular profiling measurements with diverse replication mechanisms and pathway identification from such measurements. The future advantages of our current study are at least two fold. First, our study provides a strong motivation for exploiting replicated complete and incomplete molecular profiling measurements with both blind and informed replication mechanisms in general bioinformatics frameworks. A key problem that may arise in such cases is how to accommodate replicated measurements in different pattern analysis approaches. Correlation-based patten discovery considered in this dissertation is one at-

tempt in this direction. Second, the correlation estimators developed in this dissertation may have a significant impact on the performance of other supervised and unsupervised learning approaches for pathway identification, which rely on an accurate estimate of the population correlation structure. In this dissertation, we focussed on correlation based gene clustering, which is one of the possible ways to identify pathway components from large-scale molecular profiling data. Depending on a problem scenario, other approaches, such as linear and quadratic discriminant analysis [49], co-expression networking [20, 155] and matrix factorization [17, 72], can be adapted to achieve this goal. In [154], for instance, the module discovery problem has been addressed by combining the estimation of correlation structure with matrix factorization. Our current study can be readily used in such frameworks to further accommodate replicated complete and incomplete molecular profiling measurements with diverse replication mechanisms.

*Gene set based reconstruction of large co-expression networks:* Co-expression networking is frequently used in bioinformatics analyses for inferring functional associations among genes. By considering correlation as the strength of gene-gene association, the correlation estimators presented in this dissertation can be used for co-expression networking from replicated complete and incomplete measurements with blind or informed replication mechanisms. However, another popular category of co-expression networks is represented by GGMs [125, 126], which utilize the inverse of an estimated correlation structure for inferring the strength of direct associations, known as partial correlation, among genes. Since invertibility is an issue in large p small N scenarios, shrinkage approaches [126] are typically employed for inferring the gene associations. However, these approaches require an accurate estimation of the shrinkage intensity from data, which may be problematic in the case of small sample size. Our future studies will focus on developing dimension reduction methodologies for inferring gene association networks from smaller GGMs corresponding to pathway components. In the inference of smaller networks (partial correlation matrices), a related application of our work could be to utilize the correlation estimators developed in Chapters 2-4, which can

accommodate replicated complete and incomplete measurements corresponding to a gene set with diverse replication mechanisms.

*Establishing gene set based frameworks for the Bayesian network methods:* Bayesian network methods are widely used in the inference of directed networks. However, they suffer from several issues including high computational cost, restriction of the acyclic nature of underlying network and the inference of statistical causal interactions as opposed to higher order interactions. Gene set based approaches discussed in Chapters 5-6 offer a new research direction for the structural inference of directed network topologies. Due to their inherent flexibility, our approaches can not only be extended to the frameworks used in the inference of bayesian networks, they can be more advantageous in terms of computational load and simpler methodologies in capturing higher order interaction mechanisms. We discuss some of the points below.

A well-known limitation in using bayesian network methods is huge computational load associated with the inference procedure. In many formulations, inferring a Bayesian network is an NP-hard problem, regardless of data size [24]. For example, the number of different structures for a Bayesian network with $n$ nodes, is given by the recursive formula

$$s(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} s(n-i) = n^{2^{\mathrm{O}(n)}} \tag{7.1}$$

[26, 122]. Since $s(n)$ grows exponentially with $n$, learning the network structure by exhaustively searching over the space of all possible structures is infeasible even when $n$ is small. Therefore, a tractable inference of Bayesian network relies on heuristic search algorithms such as K2 [26], MCMC [100], simulated annealing [69] and others [43, 53]. Even in this case, a number of assumptions about the number of parents for each node, a metric used to score a structure and other parameters must be made in order to to alleviate the non-trivial computational burdens associated with bayesian network inference.

The existing frameworks for learning statistical causal interactions using bayesian net-

work methods can be translated in terms of gene set based learning of signal transduction events. As mentioned above, gene set based approaches benefit from incorporating higher order interaction mechanisms in a more natural way, allow more flexibility in accommodating prior knowledge and much reduced computational burdens in several cases. Apart from the two approaches GSGS and GSSA discussed in this dissertation, we illustrated in Section 6.10 that gene set based learning of signaling pathways could also be extended to the framework of GA. We observed from the computational complexity analysis performed in Chapter 6 that gene set based GSSA approach benefitted from a manageable computational complexity and significantly better performance than traditional Bayesian network methods. In particular, we observed a much reduced computational load, both in terms of time and memory requirements, compared with sampling based MH algorithm (see Appendix A.11). The complexity of MH is often unmanageable due a large number of neighboring structures of a sampled network.

Another major limitation associated with the inference of bayesian networks is the acyclic nature of underlying topology. This limitation prohibits the inclusion of feedback effects in the network structure, which are a common biological feature. Gene set based approaches do not put such restrictions on the network structure. Although an individual gene set is viewed as a loop-free Markov chain in these approaches, the structure inferred by combining overlapping Markov chains is capable of accommodating loops. Use of the first order Markov chains in unrolling a complicated network structure into smaller building blocks also leads to a simpler computational framework.

It is clear from the above discussion that (1) gene set based approaches are able to capture the signal transduction mechanisms characterized by Markov chains in the unrolled network and (2) they allow the inference of cycles present in the network. A more important observation is that both (1) and (2), i.e. unrolling a network and inferring a network with cycles, take place in a single time slice. Since bayesian networks do not incorporate cyclic behavior in a network, they are unfolded in time for inferring cyclic mechanisms and

are referred to as dynamic bayesian networks. Dynamic bayesian network models, although useful, suffer from an inevitable increase in the model size, large computational time and memory requirements. Due to the advantages mentioned in (1) and (2) together with reduced computational load, gene set based approaches may have a significant impact in the discovery of more complex interactions from time series data. For example, our gene set based methodologies could be extended to the setting of dynamic bayesian networks for capturing the temporal associations among gene sets. This study may be useful in the discovery of non-linear patterns within a network, such as the ones obtained by connecting a gene set in a time slice to a number of parent gene sets in the previous slice.

Finally, in the inference of Bayesian networks, it is possible to learn a group of equivalent network structures representing the same joint probability distribution with the same conditional dependence and independence relations but which differ in the direction of some edges. This clearly presents an obstacle in learning the true topology. In gene set approaches, such a situation will occur when either some or all genes in an IFGS have no overlapping with the remaining IFGSs in the compendium or genes in an IFGS overlap with some of the IFGSs in the compendium, however, this overlapping is very poor. In such cases, different gene orderings of an IFGS will be equally likely. However, the above situations are less likely to hold good in real-world scenarios because (1) the number of genes in a signaling pathway is often in few hundreds. For instance, the maximum number of genes in non-metabolic signaling pathways in the KEGG database is below 400 and (2) due to increasing database knowledge, it may not be difficult to obtain a few hundred samples related to well-known diseases. (1) and (2) will together lead to the discovery of overlapping IFGSs.

It is clear from the above discussion that gene set based approaches hold strong promises in the structural study of complex signal pathways. We believe that a transformation from the existing Bayesian network methods to gene set based frameworks is necessary for broadening the scope from focusing only on pairwise interactions to the more general signal cascading events.

*Seamless integration of pathway identification and structure inference:* Identification of pathway components and structural inference of a pathway component are problems of independent interest in the field of computational systems biology. Therefore, it is necessary to develop an automatic framework that integrates the two components in one place. For example, the pathway components derived from molecular profiling data can be first utilized to construct a large-scale network by inferring signaling pathway structures corresponding to each component. The approaches GSGS, GSSA and GSGA discussed in this dissertation can be used to infer individual signaling pathway structures. In the second step, network modules can be identified by an application of community detection or network clustering algorithms on the network constructed in the previous step. The set of modules found in the second step can be used to iteratively update the set of pathway components in the previous step and *vice versa*. We aim to sufficiently exploit molecular profiling data available from diverse sources as well as prior knowledge from existing pathway databases. Overall, our focus will be to develop an easy-to-use computational and visualization tool for a seamless integration of pathway identification and structural inference of pathway components using large-scale molecular profiling data. It is the hope that multivariate models and algorithms presented in this dissertation will open a new avenue for the novel discovery of signaling pathways and their underlying mechanisms.

# Bibliography


[1] Acharya L and Zhu D. Estimating an optimal correlation structure from replicated molecular profiling data using finite mixture models. In the Procedings of *IEEE Internation Conference on Machine Learning and Applications*, 119-124, 2009.

[2] Acharya L, Judeh T, Duan Z, Rabbat M and Zhu D. GSGS: A computaional framework to reconstruct signaling pathways from gene sets, Accepted to appear in *IEEE/ACM Trans Comput Biol Bioinform.*. (Preprint *arXiv:1101.3983v3*).

[3] Acharya L, Judeh T and Zhu D. Optimal structural inference of signaling pathways from overlapping and unordered gene sets, Submitted to *Bioinformatics*, Revised August 2011.

[4] Acharya L and Zhu D. Multivariate models and algorithms for learning correlation structures from replicated molecular profiling data, *Advanced Biomedical Engineering*, Gaetano D. Gargiulo, Co-editor: Alistair McEwan (Ed.), InTech 2011.

[5] Acharya L, Judeh T and Zhu D. A survey of computational approaches to biological network reconstruction and partition, Submitted to *Wiley Publisher*, Revised April 2011.

[6] Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P. Mol Biol Cell, 4th edition, *Garland Science*, 2002.

[7] Altay G and Emmert-Streib F. Revealing differences in gene network inference algorithms on the network-level by ensemble methods, *Bioinformatics*, 26(14), 1738-1744, 2010.

[8] Altay G and Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks, *BMC Syst Biol.*, 4:132, 2010.

[9] Altay G and Emmert-Streib F. Structural influence of gene networks on their inference: analysis of C3NET, *Biol Direct*, 6:31, 2011.

[10] Altman N. Replication, variation and normalisation in microarray experiments, *Appl Bioinformatics*, 4(1):33-44, 2005.

[11] Anderson TW. An introduction to mutilvariate statistical analysis, *New York: Wiley*, 1958.

[12] Bader GD, Cary MP and Sander C. Pathguide: a pathway resource list, *Nucleic Acids Res.*, 34(Database issue):D504-506, 2006.

[13] Baker D. LVB: parsimony and simulated annealing in the search for phylogenetic trees, *Bioinformatics* 20(2), 274-275, 2004.

[14] Baldi E, Luconi M, Muratori M, Marchiani S, Tamburrino L and Forti G. Nongenomic activation of spermatozoa by steroid hormones: facts and fictions, *Mol Cell Endocrinol.*, 308(1-2), 39-46, 2009.

[15] Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A and Soboleva A. NCBI GEO: archive for functional genomics data sets-10 years on, *Nucleic Acids Res.*, 39(Database issue):D1005-1010, 2011.

[16] Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R and Califano A. Reverse engineering of regulatory networks in human B cells, *Nat Genet.*, 37:382-390, 2005.

[17] Ben-Hur A and Guyon I. Detecting stable clusters using principal component analysis in methods in molecular biology, *Humana Press*, 159–182, 2003.

[18] Boscolo R, Liao J and Roychowdhury VP. An Information Theoretic Exploratory Method for Learning Patterns of Conditional Gene Coexpression from Microarray Data, *IEEE/ACM Trans Comput Biol Bioinform.*, 15-24, 2008.

[19] Butte AJ and Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, *Pac. Symp. Biocomput.* 5, 415-426, 2000.

[20] Butte AS and Kohane IS. Relevance networks: a first step toward finding genetic regulatory networks within microarray data in *The Analysis of Gene Expression Data* (eds Parmigiani, G., Garett, E.S., Irizarry, R.A. and Zeger, S.L.), Springer, New York, 428-446, 2003.

[21] Casella G and Berger RL. Statistical inference, *Duxbury Advanced Series*, 1990.

[22] Chandarlapaty S, Sawai A, Scaltriti M, Rodrik-Outmezguine V, Grbovic-Huezo O, Serra V, Majumder PK, Baselga J and Rosen N. AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity, *Cancer Cell*, 19(1):58-71, 2011.

[23] Chen CM, Lee C, Chuang CL, Wang CC and Shieh GS. Inferring genetic interactions via a nonlinear model and an optimization algorithm, *BMC Syst Biol.*, 4(16), 2010.

[24] Chickering DM. Optimal structure identification with greedy search, *J Mach Learn Res.*, 3, 507-554, 2002.

[25] Chong EKP and Żak SH. An Introduction to Optimization, 3rd edition *John Wiley & Sons*, 2008.

[26] Cooper GF and Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data, *Machine Learning*, 9(4), 309-347, 1992.

[27] Cui XQ and Churchill GA. Statistical tests for differential expression in cDNA microarray experiments, *Genome Biol.*, 4:201, 2003.

[28] Davis L, (Ed.). Genetic Algorithms and Simulated Annealing, *Research Notes in Artificial Intelligence, London: Pitman*, 1987.

[29] Dempster AP, Laird NM and Rubin DB. Maximum Likelihood from incomplete data via the EM algorithm, *J R Stat Soc Series B*, 39(1):1-38, 1977.

[30] Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA. DAVID: Database for annotation, visualization and integrated discovery, *Genome Biol.*, 4(5):P3, 2003.

[31] Dimitrova ES, Licona MP, McGee J and Laubenbacher R. Discretization of time series data, *J. Comp. Biol.*, 17(6), 853-868, 2010.

[32] Dobra A, Hans C, Jones B, Nevins JR and West M. Sparse graphical models for exploring gene expression data, *J. Multiv. Anal.*, 90, 196-212, 2004.

[33] Drăghici S, Khatri P, Martins RP, Ostermeier GC and Krawetz SA. Global functional profiling of gene expression, *Genomics.*, 81(2):98-104, 2003.

[34] Eisen M, Spellman P, Brown PO and Botstein D. Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA*, 95:14863-14868, 1998.

[35] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ and Gardner TS. Large-Scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, *PLoS Biol* 5(1):e8, 2007.

[36] Fraley C and Raftery AE. Model-based clustering, discriminant analysis and density estimation, *J Am Stat Assoc, 97*, 611-631, 2002.

[37] Friedman N, Linial M, Nachman I and Peer D. Using Bayesian networks to analyze expression data, *J Comput Biol.*, 7, 601-620, 2000.

[38] Gardner TS, di Bernardo D, Lorenz D and Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling, *Science*, 301, 5629, 102-105, 2003.

[39] Gatza ML, Lucas JE, Barry WT, Kim JW, Wang Q, Crawford MD, Datto MB, Kelley M, Mathey-Prevot B, Potti A and Nevins JR. A pathway-based classification of human breast cancer, *Proc Natl Acad Sci USA*,107(15):6994-6999, 2010.

[40] Gelman A, Carlin JB, Stern HS and Rubin DB: Bayesian Data Analysis. *Chapman & Hall*, 2nd Edition, 2003.

[41] Givens GH and Hoeting JA. Computational Statistics, *Wiley Series in Proabbility and Statistics*, 2005.

[42] Glaab E, Baudot A, Krasnogor N and Valencia A. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9): 1271-1272, 2010.

[43] Glover F. Tabu Search - Part I, *ORSA J Comp.*, Vol. 1, No. 3, 190-206, 1989.

[44] Goldstein L and Waterman M. Neighborhood size in the simulated annealing algorithm, *Amer. J. Math. Management Sci.*, 8, 3-4, 1988.

[45] Gonzalez OR, Küper C, Jung K, Naval PC Jr and Mendoza E. Parameter estimation using Simulated Annealing for S-system models of biochemical networks, *Bioinformatics*, 23(4): 480-486, 2007.

[46] Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J, Doucet D, Milewski M, Yang R, Siegmund C, Haas J, Zhou L, Oliphant A, Fan JB, Barnard S and Chee MS. Decoding randomly ordered DNA arrays, *Genome Res.* 14, 870-877, 2004.

[47] Hajek B. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, Vol 13, No. 2, 311-329, 1988.

[48] Hartigan JA and Wong MA. A k-means clustering algorithm, *Applied Stat.*, 28, 100-108, 1979.

[49] Hastie T, Tibshirani R and Friedman J. The Elements of Statistical Learning: Prediction, Inference and Data Mining, *Springer-Verlag*, New York, 2009.

[50] Hathaway RJ. A constrained formulation of maximum-likelihood estimation for normal mixture distributions, *Ann. Statist. 13*, 795–800, 1985.

[51] Heckerman D, Geiger D and Chickering, M. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20, 197-243, 1995.

[52] Heyer LJ, Kruglyak S and Yooseph S (1999). Exploring expression data: identification and analysis of coexpressed genes, *Genome Res.*, 9, 1106-1115.

[53] Holland JH. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, *MIT Press*, Cambridge, MA, 1992.

[54] Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Krieger CJ, Livstone MS, Miyasato SR, Nash RS, Oughtred R, Skrzypek MS, Weng S, Wong ED, Zhu KK, Dolinski K, Botstein D and Cherry JM. Gene Ontology annotations at SGD: new data sources and annotation methods, *Nucleic Acids Res.*, 36(Database issue):D577-581, 2008.

[55] de Hoon MJL, Imoto S, Nolan J and Miyano, S: Open source clustering software, *Bioinformatics*, 20(9):1453-1454, 2004.

[56] Huang S. Gene expression profiling, genetic networks and cellular states: An integrating concept for tumorigenesis and drug discovery. *J Mol Med.*, 77, 469-480, 1999.

[57] Huang DW, Sherman BT and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat Protoc.*, 4(1), 44-57, 2009.

[58] Hubbell E, Liu WM, and Mei R. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585-1592, 2001.

[59] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H and He YD. Functional discovery via a compendium of expression profiles, *Cell*, 102:109-126, 2000.

[60] Ideker T, Thorsson V, Siegel AF and Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data, *J. Comput. Biol*, 7:805-817, 2000.

[61] Ingrassia S. A likelihood-based constrained algorithm for multivariate normal mixture models, *Stat Methods Appt. 13*, 151-166, 2004.

[62] Ingrassia S and Rocci R. Constrained monotone EM algorithms for the finite mixtures of multivariate Gaussians, *Comput Stat Data Anal. 51*, 5399-5351, 2007.

[63] Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A and di Bernardo D. Discovery of drug mode of action and drug repositioning from transcriptional responses, *Proc Natl Acad Sci USA*, 107(33), 14621-14626, 2010.

[64] Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4:249-264, 2003.

[65] Jardine N and Sibson R. Mathematical Taxonomy, *John Wiley and Sons*, 1971.

[66] Kaderali L, Dazert E, Zeuge U, Frese M and Bartenschlager R. Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks, *Bioinformatics*, 25(17): 2229-2235, 2009.

[67] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M and Hirakawa M. From genomics to chemical genomics: new developments in KEGG, *Nucleic Acids Res.*, 34 (Database issue):D354-357, 2006.

[68] Kanehisa M, Goto S, Furumichi M, Tanabe M and Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res.*, 38, D355-360, 2010.

[69] Kernighan BW and Lin S. An efficient heuristic procedure for partitioning graphs, *Bell Systen Technical Journal*, 49, 291-307, 1970.

[70] Kerr MK and Churchill GA. Experimental design for gene expression microarrays. *Biostatistics*, 2:183-201, 2001.

[71] Khatri P and Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics*, 21, 3587-3595, 2005.

[72] Kim PM and Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data, *Genome Res.*, 13(7), 1706-1718, 2003.

[73] Kim Y, Son S and Jeong H. Finding communities in directed networks, *Phys. Rev. E*, 81(1):016103, 2010.

[74] Kirkpatrick S, Gelatt CD Jr and Vecchi MP. Optimization by simulated annealing, *Science*, 220, 671-680, 1983.

[75] Kishino H and Waddell PJ. Correspondence analysis of genes and tissue types and finding genetic links from microarray data, *Genome Informatics*, 11, 83-95, 2000.

[76] Kubica J, Moore A, Cohn D and Schneider J. cGraph: A fast graphbased method for link analysis and queries, *Proceedings of IJCAI Text-Mining and Link-Analysis Workshop*, Acapulco, Mexico, 2003.

[77] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES and Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, *Science*, 313(5795), 1929-1935, 2006.

[78] Leicht E and Newman MEJ. Community structure in directed networks. *Physical Review Lett.*, 100(11):118703, 2008.

[79] Leary AF, Sirohi B, Johnston SR. Clinical trials update: endocrine and biological therapy combinations in the treatment of breast cancer, *Breast Cancer Res.*, 9(5):112, 2007.

[80] Leary AF, Drury S, Detre S, Pancholi S, Lykkesfeldt AE, Martin LA, Dowsett M, Johnston SR. Lapatinib restores hormone sensitivity with differential effects on estrogen receptor signaling in cell models of human epidermal growth factor receptor 2-negative breast cancer with acquired endocrine resistance. *Clin Cancer Res.*, 16(5):1486-1497, 2010.

[81] Li C and Wong WH. Model-based analysis of oligonucleotide arrays: expression score computation and outlier detection, *Proc Natl Acad Sci USA*, 98:31-36, 2001.

[82] Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat Biotech.*, 14, 1675-1680, 1996.

[83] Lurje G and Lenz HJ. EGFR signaling and drug discovery, *Oncology*, 77(6), 400-410, 2009.

[84] Luxburg U. A tutorial on spectral clustering in *Statistics and Computing*, 17(4), 395-416, 2007.

[85] Mahata P. Exploratory consensus of hierarchical clusterings for Melanoma and Breast Cancer, *IEEE/ACM Trans Comput Biol Bioinform.*, Jan-Mar 7(1):138-152, 2010.

[86] Marbach D, Schaffter T, Mattiussi C and Floreano D. Generating realistic in silico gene networks for performance assessment of reverse engineering methods, *J Comput Biol.*, 16(2), 229-239, 2009.

[87] Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D and Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference, *Proc Natl Acad Sci USA*, 107(14), 6286-6291, 2010.

[88] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R and Califano A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context, *BMC Bioinform.*, Suppl 1, S7, 2006.

[89] Markowetz F, Kostka D, Troyanskaya OG, Spang R. Nested effects models for high-dimensional phenotyping screens, *Bioinformatics*, 23(13), i305-312, 2007.

[90] McLachlan GJ and Peel D. Finite Mixture Models, *Wiley series in Probability and Mathematical Statistics, Applied Probability and Statistics Section, New York, John Wiley & Sons*, 2000.

[91] McLachlan GJ and Peel D. On computational aspects of clustering via mixtures of normal and t-components, *Proceedings of the American Statistical Association (Bayesian Statistical Science Section), Indianapolis, Virginia*, 2000.

[92] Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, Al-Shahrour F and Dopazo J. Gene set-based analysis of polymorphisms: finding pathways or

biological processes associated to traits in genome-wide association studies, *Nucleic Acids Res.*, 37, 340-344, 2009.

[93] Medvedovic M and Sivaganesan S. Bayesian infinite mixture model based clustering of gene expression profiles, *Bioinformatics*, 18, 1194-1206, 2002.

[94] Medvedovic M, Yeung KY and Bumgarner RE. Bayesian mixtures for clustering replicated microarray data, *Bioinformatics*, 20, 222-1232, 2004.

[95] Mendes P. Framework for Comparative Assessment of Parameter Estimation and Inference Methods in Systems Biology, *Learning and Inference in Computational Systems Biology* (Lawrence, N.D., Girolami, M., Rattray, M., Sanguinetti, G. eds.), MIT Press, Cambridge, MA, 33-58, 2009.

[96] Meyer PE, Kontos K and Bontempi G. Information-theoretic inference of large transcriptional regulatory networks, *EUROSIP J Bioinform Syst Biol.*, 79879, 2007.

[97] Meyer PE, Lafitte F and Bontempi. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information, *BMC Bioinform.*, 9:461, 2008.

[98] Mitchell M. An Introduction to Genetic Algorithms, *MIT Press, Cambridge, MA*, 1996.

[99] Mortazavi A, Williams B, McCue K, Schaeffer L and Wold, B. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq, *Nat Method*, 5:621-628, 2008.

[100] Murphy K. Active learning of causal bayes net structure, Technical Report, UC Berkeley, 2001.

[101] Murphy K. The Bayes net toolbox for MATLAB, *Computing Science and Statistics: Proceedings of Interface*, 33, 2001.

[102] Murphy K. Bayes Net Toolbox v5 for MATLAB. Cambridge, MA, MIT AI Lab, 2003.

[103] Nasser S, Cunliffe HE, Black MA and Kim S. Context-specific gene regulatory networks subdivide intrinsic subtypes of breast cancer, *BMC Bioinform.*, 12 Suppl 2:S3, 2011.

[104] Navolanic PM, Steelman LS and McCubrey JA. EGFR family signaling and its association with breast cancer development and resistance to chemotherapy (Review), *Int J Oncol.*, 22(2):237-252, 2003.

[105] Newman MEJ and Girvan M. Finding and evaluating community structure in networks, *Phys. Rev. E*, 69(2):026113, 2004.

[106] Newman MEJ. Modularity and community structure in networks, *Proc Natl Acad Sci USA*, 103(23), 8577-8582, 2006.

[107] Newman MEJ and Leicht E. Mixture models and exploratory analysis in networks, *Proc Natl Acad Sci USA*, 104(23), 9564-9569, 2007.

[108] Nguyen DV and Rocke DM. Multi-class cancer classification via partial least squares using gene expression profiles, *Bioinformatics*, 18(9), 1216-1226, 2002.

[109] Olayioye MA. Update on HER-2 as a target for cancer therapy: intracellular signaling pathways of ErbB2/HER-2 and family members, *Breast Cancer Res.*, 3(6), 385-389, 2001.

[110] Osborne CK, Neven P, Dirix LY, Mackey JR, Robert J, Underhill C, Schiff R, Gutierrez C, Migliaccio I, Anagnostou VK, Rimm DL, Magill P and Sellers M. Gefitinib or Placebo in Combination with Tamoxifen in Patients with Hormone Receptor-Positive Metastatic Breast Cancer: A Randomized Phase II Study. *Clin Cancer Res.*, 17(5):1147-1159, 2011.

[111] Palla G, Derenyi I, Farkas I and Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, 435(7043), 814-818, 2005.

[112] Pang H, Lin A, Holford M, Enerson BE, Lu B, Lawton MP, Floyd E and Zhao H. Pathway analysis using random forests classification and regression, *Bioinformatics*, 22, 2028-2036, 2006.

[113] Pang H and Zhao H. Building pathway clusters from Random Forests classification using class votes, *BMC Bioinform.*, 9(87), 2008.

[114] Park CY, Hess DC, Huttenhower C and Troyanskaya OG. Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components, *PLoS Comp Bio.*, 6(11):e1001009, 2010.

[115] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E, Kurbatova N, Lukk M, Malone J, Mani R, Pilicheva E, Rustici G, Sharma A, Williams E, Adamusiak T, Brandizi M, Sklyar N, Brazma A. ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments, *Nucleic Acids Res.*, 39(Database issue):D1002-1004, 2011.

[116] Pehkonen P, Wong G, Törönen P. Heuristic Bayesian segmentation for discovery of co-expressed genes within genomic regions, *IEEE/ACM Trans Comput Biol Bioinform*, 7(1):37-49, 2010.

[117] Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G and Stolovitzky G. Towards a rigorous assessment of systems biology models: the DREAM3 challenges, *PLoS ONE*, 5(2):e9202, 2010.

[118] Rabbat MG, Treichler JR, Wood SL and Larimore MG: Understanding the topology of a telephone network via internallysensed network tomography, *Proc. IEEE International Confernece on Acoustics, Speech, and Signal Processing*, 3, Philadelphia, PA, 977980, 2005.

[119] Rabbat MG, Figueiredo MAT and Nowak RD: Network inference from co-occurrences. *IEEE Trans Inf Theory*, 54(9), 4053-4068, 2008.

[120] Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP and Young RA. Genome-wide location and function of DNA binding proteins, *Science*, 290(5500):2306-2309, 2000.

[121] Richards AJ, Muller B, Shotwell M, Cowart LA, Baerbel R and Lu X: Assessing the functional coherence of gene sets with metrics based on the Gene Ontology graph, *Bioinformatics*, 26(12), i79-87, 2010.

[122] Robinson RW. Counting unlabeled acyclic digraphs in *Combinatorial Mathematics V* (ed Little, C.H.C.), Lecture Notes in Mathematics, 622, 28-43, Berlin, Springer, 1977.

[123] Roskoski R Jr. RAF protein-serine/threonine kinases: structure and regulation, *Biochem Biophys Res Commun.*, 399(3), 313-317, 2010.

[124] Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD and Medvedovic M. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments, *BMC Bioinform.*, 7:538, 2006.

[125] Schäfer J and Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics*, 21, 754-764, 2005.

[126] Schäfer J and Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat Appl Genet Mol Biol.*, 4, Article 32, 2005.

[127] Schwartz G. Estimating the dimension of a model, *Ann Stat.*, 6(2), 461-464, 1978.

[128] Schena M, Shalon D, Davis RW and Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270 (5235), 368-371, 1995.

[129] Schiff R, Massarweh SA, Shou J, Bharwani L, Mohsin SK and Osborne CK. Cross-talk between estrogen receptor and growth factor pathways as a molecular target for overcoming endocrine resistance, *Clinical Cancer Res.*, 10, 331S-336S, 2004.

[130] Segal E, Shapira M, Regev A, Peer D, Botstein D, Koller D and Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat. Genet.*, 34, 166-176, 2003.

[131] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.*, 13(11), 2498-2504, 2003.

[132] Shedden K and Taylor J. Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas, *Methods of Microarray Data Analysis IV* (Jennifer Shoemaker ed.), Kluwer, 2004.

[133] Shendure J, Mitra RD, Varma C and Church GM. Advanced sequencing technologies: methods and goals, *Nat Rev Genet.*, 5(5), 335-344, 2004.

[134] Shendure J and Ji H. Next-generation DNA sequencing, *Nat. Biotech.*, 26, 1135-1145, 2008.

[135] Shmulevich I, Dougherty ER, Kim S and Zhang W. Probabilistic Boolean Networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, 18(2), 261-274, 2002.

[136] Shmulevich I, Gluhovsky I, Hashimoto R, Dougherty ER and Zhang W. Probabilistic Boolean Networks: A Rule-based Uncertainty Model for Gene Regulatory Networks, *Comp Funct Genomics.*, 4(6), 601-608, 2003.

[137] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes

of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell.*, 9(12):3273-3297, 1998.

[138] Staal FJ, van der Burg M, Wessels LF, Barendregt BH, Baert MR, van den Burg CM, van Huffel C, Langerak AW, van der Velden VH, Reinders MJ, van Dongen JJ. DNA microarrays for comparison of gene expression profiles between diagnosis and relapse in precursor-B acute lymphoblastic leukemia: choice of technique and purification influence the identification of potential diagnostic markers, *Leukemia*, 17(7):1324-1332, 2003.

[139] Stolovitzky G, Prill RJ and Califano A. Lessons from the DREAM2 challenges in *Annals of the New York Academy of Sciences* (eds Stolovitzky, G., Kahlem, P., Califano, A.), 1158, 159-195, 2009.

[140] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci USA*, 102, 15545-15550, 2005.

[141] Taieb FE, Gross SD, Lewellyn AL and Maller JL. Activation of the anaphase-promoting complex and degradation of cyclin B is not required for progression from Meiosis I to II in Xenopus oocytes, *Curr. Biol.*, 11(7):508-513, 2001.

[142] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES and Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc Natl Acad Sci USA*, 96, 2907-2912, 1999.

[143] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis, *Bioinformatics*, 25(1), 75-82, 2009.

[144] Tegner J, Yeung MKS, Hasty J and Collins JJ. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling, *Proc Natl Acad Sci USA*, 100(10), 5944-5949, 2003.

[145] Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS and Park PJ. Discovering statistically significant pathways in expression profiling studies, *Proc Natl Acad Sci USA*, 102(38), 13544-13549, 2005.

[146] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17, 520525, 2001.

[147] Tusher VG, Tibshirani R and Chu G. Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci USA*, 98(9):5116-21, 2001.

[148] Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D and Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, *Bioinformatics*, 26(12), i237-245, 2010.

[149] Wang K, Nemenman I, Banerjee N, Margolin AA and Califano A. Genome-wide discovery of modulators of transcriptional interactions in human B lymphocytes, *RECOMB'06*, 348-362, 2006.

[150] Wu Z and Irizarry RA. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J Comput Biol.*, 12:882-893, 2005.

[151] Yao J, Chang C, Salmi ML, Hung YS, Loraine A, Roux SJ. Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient, *BMC Bioinform.*, 9:288, 2008.

[152] Yeung KY, Medvedovic M and Bumgarner RE. Clustering gene-expression data with repeated measurements, *Genome Biol.*, 4(5):R34, 2003.

[153] Yeung KY and Bumgarner R. Multi-class classification of microarray data with repeated measurements: application to cancer, *Genome Biol.*, 6(405), 2005.

[154] Yang X, Zhou Y, Jin R and Chan C. Reconstruct modular phenotype-specific gene networks by knowledge-driven matrix factorization, *Bioinformatics*, 25(17):2236-2243, 2009.

[155] Zhu D, Hero AO, Qin ZS and Swaroop A. High throughput screening of co-expressed gene pairs with controlled False Discovery Rate (FDR) and Minimum Acceptable Strength (MAS), *J Comput Biol.*, 12(7), 1027-1043, 2005.

[156] Zhu D, Hero AO, Cheng H, Khanna R and Swaroop A. Network constrained clustering for gene microarray data, *Bioinformatics*, 21(21):4014-4020, 2005.

[157] Zhu D, Rabbat MG, Hero AO, Nowak R and Figueirado MAG. *De Novo* Reconstructing Signaling Pathways from Multiple Data Sources, In a chapter of the book *New Research in Signaling Transduction*, Nova Publisher, New York, 2006.

[158] Zhu D, Li Y and Li H. Multivariate correlation estimator for inferring functional relationships from replicated genome-wide data, *Bioinformatics, 23(17)*, 2298-2305, 2007.

[159] Zhu D and Hero AO. Bayesian hierarchical model for large covariance matrix estimation, *J Comput Biol.*, 14(10), 1311-1326, 2007.

[160] Zhu D, Dequéant ML and Li H. Comparative analysis of distance based clustering methods in *Analysis of microarray data: A network based approach* (eds Emmert-Streib, F. and Dehmer, M.), Wiley-VCH, Weinheim, Germany, 2007.

[161] Zhu D, Acharya L and Zhang H. A generalized multivariate approach to pattern discovery from replicated and incomplete genome-wide measurements, *IEEE/ACM Trans Comput Biol Bioinform.*, 8(5):1153-1169, 2011.

# Chapter A

# Appendix

## A.1 Derivation of the MLEs $\hat{\mu}^I$ and $\hat{\Sigma}^I$

The likelihood function of a $(m_1 + m_2)$-variate normal family is given as

$$L(\mu^I, \Sigma^I) = \frac{1}{(2\pi)^{n(m_1+m_2)/2}|\Sigma^I|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(Z_j-\mu^I)^T{\Sigma^I}^{-1}(Z_j-\mu^I)}. \tag{A.1}$$

The log-likelihood function is therefore

$$\mathcal{L}(\mu^I, \Sigma^I) = nlog|\Sigma^I| + \sum_{j=1}^{n}(Z_j - \mu^I)^T{\Sigma^I}^{-1}(Z_j - \mu^I). \tag{A.2}$$

The equation

$$\frac{\partial\mathcal{L}(\mu^I, \Sigma^I)}{\partial\mu_{g_1}^{j_{m_1}}} \;=\; 0,\; 1 \le j_{m_1} \le J_{m_1} \tag{A.3}$$

leads to

$${\Sigma^I}^{-1}\sum_{j=1}^{n}(Z_j - \mu^I)V = 0. \tag{A.4}$$

where $V$ is a $1 \times (m_1 + m_2)$ vector with $V_t = 1$, for $\sum_{l=1}^{j-1} I_{m_1}^l < t \le \sum_{l=1}^{j} I_{m_1}^l$, and 0 otherwise. Since

$$\sum_{j=1}^{n} tr((Z_j - \mu^I)V) = \sum_{j=1}^{n} V(Z_j - \mu^I) = 0$$

we have

$$\hat{\mu}_{g_1}^{j_{m_1}} = \frac{1}{I_{m_1}^j n}\sum_{k=1}^{n}\sum_{i=\sum_{l=1}^{j} I_{m_1}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_1}^l} g_{ik}^1,\; 1 \le j_{m_1} \le J_{m_1}. \tag{A.5}$$

Similarly,

$$\hat{\mu}_{g_2}^{j_{m_2}} = \frac{1}{I_{m_2}^{j} n} \sum_{k=1}^{n} \sum_{i=\sum_{l=1}^{j} I_{m_2}^{l-1}+1}^{\sum_{l=1}^{j} I_{m_2}^{l}} g_{ik}^2, \ 1 \leq j_{m_2} \leq J_{m_2} \tag{A.6}$$

Thus $\mu^I$ is estimated as

$$\hat{\mu}^I = \left(\hat{\mu}_{g_1}^1, \ldots, \hat{\mu}_{g_1}^1, \ldots, \hat{\mu}_{g_1}^{J_{m_1}}, \ldots, \hat{\mu}_{g_1}^{J_{m_1}}, \hat{\mu}_{g_2}^1, \ldots, \hat{\mu}_{g_2}^1, \ldots, \hat{\mu}_{g_2}^{J_{m_2}}, \ldots, \hat{\mu}_{g_2}^{J_{m_2}}\right)^T \tag{A.7}$$

To find $\hat{\Sigma}^I$, let us consider:

$$l(\mu^I, \Sigma^I) = \frac{n(m_1+m_2)}{2} ln 2\pi - \frac{n}{2} ln|\Sigma^I| - \frac{1}{2}\sum_{j=1}^{n} (Z_j - \mu^I)^T \Sigma^{-1} (Z_j - \mu^I)$$

$$= \frac{n(m_1+m_2)}{2} ln 2\pi - \frac{n}{2} ln|\Sigma^I| - \frac{1}{2}\sum_{j=1}^{n} tr(Z_j - \mu^I)^T \Sigma^{-1} (Z_j - \mu^I)$$

$$= \frac{n(m_1+m_2)}{2} ln 2\pi - \frac{n}{2} ln|\Sigma^I| - \frac{1}{2}\sum_{j=1}^{n} tr\Sigma^{-1} (Z_j - \mu^I)(Z_j - \mu^I)^T$$

Thus,

$$\frac{\partial l(\mu^I, \Sigma^I)}{\partial \Sigma^I} = -\frac{n}{2}\frac{\partial ln|\Sigma^I|}{\partial \Sigma^I} - \frac{1}{2}\frac{\partial}{\partial \Sigma^I}\sum_{j=1}^{n} tr\Sigma^{I^{-1}} (Z_j - \mu^I)(Z_j - \mu^I)^T$$

$$= -\frac{n}{2}\Sigma^{I^{-1}} + \frac{1}{2}\Sigma^{I^{-1}}\sum_{j=1}^{n} (Z_j - \mu^I)(Z_j - \mu^I)^T \Sigma^{I^{-1}}.$$

Now,

$$\frac{\partial l(\mu^I, \Sigma^I)}{\partial \Sigma^I} = 0$$

leads to

$$\hat{\Sigma}^I = \frac{1}{n}\sum_{j=1}^{n} \begin{pmatrix} (Z_j^{[1]} - \hat{\mu}^{[1]})(Z_j^{[1]} - \hat{\mu}^{[1]})^T & (Z_j^{[1]} - \hat{\mu}^{[1]})(Z_j^{[2]} - \hat{\mu}^{[2]})^T \\ (Z_j^{[2]} - \hat{\mu}^{[2]})(Z_j^{[1]} - \hat{\mu}^{[1]})^T & (Z_j^{[2]} - \hat{\mu}^{[2]})(Z_j^{[2]} - \hat{\mu}^{[2]})^T \end{pmatrix} \tag{A.8}$$

Equation A.7 and Equation A.8 give the closed form formulae for the MLE of $\mu^I$ and $\Sigma^I$.

## A.2 Summarization of Correlation

We are interested in the sum $\sum_{j=1}^{n}(Z_j^{[1]} - \mu^{I^{[1]}})(Z_j^{[2]} - \mu^{I^{[2]}})^T$. It is easy to see that the sum of elements in $(Z_j^{[1]} - \mu^{I^{[1]}})(Z_j^{[2]} - \mu^{I^{[2]}})^T$ is given by

$$(\sum_{i=1}^{m_1} g_{ij}^1 - \sum_{i=1}^{m_1} \mu_i^{I^{[1]}})(\sum_{i=1}^{m_2} g_{ij}^2 - \sum_{i=1}^{m_2} \mu_i^{I^{[2]}})^T,$$

which is equal to

$$(\sum_{i=1}^{m_1} g_{ij}^1 - m_1\mu_{g_1}^B)(\sum_{i=1}^{m_2} g_{ij}^2 - m_2\mu_{g_2}^B)$$

for the parameters $\mu_{g_1}^B$ and $\mu_{g_2}^B$ of the blind case estimator. The latter expression corresponds to the sum of elements in $(Z_j^{[1]} - \mu^{B^{[1]}})(Z_j^{[2]} - \mu^{B^{[2]}})^T$. This is true for each $j = 1, \ldots, n$. Thus, we get the same estimate of between-molecular correlation from two models, by the method of averaging the cross-diagonal blocks in the estimated correlation matrix.

## A.3 Missing Values Imputation Using K-Nearest Neighbors

The k-nearest neighbor algorithm used for missing data imputation is based on selecting neighboring genes with similar expression profiles as the gene with missing measurements. The information missing in a gene is gained from the neighboring genes where this information is present. For instance, if the expression level of a gene is missing in an experiment, the method seeks for $k$ other genes, for which the expression levels in that experiment are known. The nearness of genes is computed by using Euclidean distance as metric. The missing value is then imputed by averaging the expression levels from $k$ nearest neighbors. The algorithm has been implemented in an R package `impute` which can be installed from CRAN `http://cran.r-project.org`.

## A.4 SD-Weighted Correlation

Suppose we have $n$ samples consisting of $m_1$ replicated measurements corresponding to gene $X$ and $m_2$ replicated measurements for gene $Y$. The mean and variance of gene $X$ and $Y$ in the $j^{th}$ sample is defined as follows

Mean:

$$\overline{M}_{G_1}(j) = \sum_{i=1}^{m_1} g_{ij}^1/m_1$$

$$\overline{M}_{G_2}(j) = \sum_{i=1}^{m_2} g_{ij}^2/m_2$$

Variance:

$$S_{G_1}^2(j) = \frac{1}{m_1 - 1} \sum_{i=1}^{m_1} (g_{ij}^1 - \overline{M}_{G_1}(j))^2$$

$$S_{G_2}^2(j) = \frac{1}{m_2 - 1} \sum_{i=1}^{m_2} (g_{ij}^2 - \overline{M}_{G_2}(j))^2$$

for $j = 1, \ldots, n$. Using standard deviation(SD) as a criterion to measure error, the SD-weighted average expressions of genes $G_1$ and $G_2$ across all the samples are given by

$$\overline{M}_{G_1} = \sum_{j=1}^{n} \frac{\overline{M}_{G_1}(j)}{S_{G_1}^2(j)} \; / \; \sum_{j=1}^{n} \frac{1}{S_{G_1}^2(j)}$$

and

$$\overline{M}_{G_2} = \sum_{j=1}^{n} \frac{\overline{M}_{G_2}(j)}{S_{G_2}^2(j)} \; / \; \sum_{j=1}^{n} \frac{1}{S_{G_2}^2(j)}.$$

The SD-weighted correlation coefficient [151] is defined as

$$\rho_{G_1 G_2} = \frac{\sum_{j=1}^{n} \left( \frac{\overline{M}_{G_1}(j) - \overline{M}_{G_1}}{S_{G_1}(j)} \right) \left( \frac{\overline{M}_{G_2}(j) - \overline{M}_{G_2}}{S_{G_2}(j)} \right)}{\sqrt{\sum_{j=1}^{n} \left( \frac{\overline{M}_{G_1}(j) - \overline{M}_{G_1}}{S_{G_1}(j)} \right)^2 \left( \sum_{j=1}^{n} \frac{\overline{M}_{G_2}(j) - \overline{M}_{G_2}}{S_{G_2}(j)} \right)^2}}$$

## A.5 Description of the Bayesian Network Methods

In principle, the K2 approach [26] begins by specifying an ordering of nodes involved in the underlying network. Thus, initially each node has no parent. The algorithm incrementally assigns a parent to a node whose addition increases the score of the resulting structure the most. For the $i^{th}$ node, parents are chosen from the set of nodes with index $1, \ldots, i-1$. On the other hand, the MH algorithm [100] starts with an initial directed acyclic graph (DAG) $Gr_0$ and selects a network $Gr_1$ uniformly from the neighborhood of $Gr_0$. The neighborhood of a network $Gr$ is the collection of all DAGs which differ from $Gr$ by addition, deletion or reversal of a single edge. The algorithm accepts or rejects the move from $Gr_0$ to $Gr_1$ by computing an acceptance ratio defined in terms of marginal likelihood ratio $P(D|Gr_1)/P(D|Gr_0)$, where $D$ represents the given data. This procedure is iterated starting from the most recent network. A specified number of networks are collected after burn-in state. For scoring a structure, BNT provides Bayesian Information Criterion [127] and Bayesian score function [26], where Bayesian score function is defined for discrete measurements. Both K2 and MH have been implemented in the Bayes Net Tool Box (BNT) [101].

Here we define two Bayesian score functions *Bayesian Dirichlet (BD) score* from [51] and *K2 score* presented in [26], and *Bayesian Information Criterion (BIC score)* [127].

BD score is defined as [51]

$$P(Gr, D) = P(Gr) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ijk})}, \tag{A.9}$$

where $n$ is the number of genes, $r_i$ represents the number of states of $x_i$, $q_i = \prod_{x_j \in Pa(x_i)} r_j$, $N_{ijk}$ is the number of times $x_i$ is in $k^{th}$ state and members in $Pa(x_i)$ are in $j^{th}$ state, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$, $N_{ik} = \sum_{j=1}^{q_i} N_{ijk}$, $N'_{ijk}$ are the parameters of Dirichlet prior distribution, $P(Gr)$ stands for the prior probability of the structure $Gr$ and $\Gamma()$ represents the Gamma function.

The K2 score is given by [26]

$$P(Gr, D) = P(Gr) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{A.10}$$

We refer to [26, 51] for further readings on Bayesian score functions.

BIC score is defined as

$$Pr(Gr, D) = \log P(D|Gr, \theta^{ML}) - \frac{1}{2} \left(\sum_{i=1}^{n} (r_i - 1) q_i\right) \log N, \tag{A.11}$$

where $N$ is sample size and $\theta^{ML}$ are parameter values obtained by likelihood maximization.

## A.6 Description of the Mutual Information Methods

Description of various mutual information network inference methods used in this dissertation, available from the R packages MINET [97] and C3NET [8] is as follows. RN is based on assigning to each pair of nodes a weight equal to their mutual information. This is followed by eliminating non-significant mutual information values. Significant weights are considered as true edges. ARACNE also first estimates mutual information between each pair of nodes. It then eliminates the weakest edge among each triplet, if the difference between the two lowest weights is above a specified threshold `eps`. The CLR algorithm is an extension of relevance network. Instead of considering the mutual information $I(X_i, X_j)$ between features $X_i$ and $X_j$, it takes into account the score $\sqrt{(z_i^2 + z_j^2)}$, where $z_i = \max(0, (I(X_i, X_j) - \text{mean}(X_i))/\text{sd}(X_i))$ and $\text{mean}(X_i)$ and $\text{sd}(X_i)$ are, respectively, the mean and the standard deviation of the empirical distribution of the mutual information values $I(X_i, X_k)$, $k = 1, ..., n$. The MRNET approach uses a MRMR (Maximum Relevance Minimum Redundancy) feature selection procedure for each variable of the dataset. The MRMR method starts by selecting the variable $X_i$ having the highest mutual information with the target $Y$. In the following steps, given a set $S$ of selected variables, the criterion

updates $S$ by choosing the variable $X_k$ that maximizes $I(X_k, Y) - \text{mean}(I(X_k, X_i))$, $X_i \in S$. The weight of each pair $(X_i, X_j)$ will be the maximum score between the one computed when $X_i$ is the target and the one computed when $X_j$ is the target. The C3NET algorithm first infers a RN. It keeps all maximum valued mutual information values for each row in the matrix and sets the rest of the elements in the matrix zero.

## A.7 Generation of All Linear Paths from a Network

**Algorithm A.1** Network2GeneSets

```
1: Input: A directed acyclic graph with n nodes
2: Output: All IFGSs
3: for i = 1, ..., n do
4:    if node i has no children then
5:       continue
6:    else
7:       add to Queue Q and the Linked List L all the directed pairs consisting of i and a child of
         i
8:       while Q is not empty do
9:          Pop an information flow P from Q
10:         if the last node in P, say k, has no children then
11:            continue
12:         end if
13:         add to Q and L, all information flows obtained by appending each child of k to P
14:      end while
15:   end if
16: end for
17: Prune information flows in L of length 2 (prior knowledge)
18: Randomly permute orders of information flows in L and order of genes in each information flow
19: Return all IFGSs of length ≥ 3.
```

## A.8 Generation of BFS Paths from a Network

After extracting networks from the KEGG database, IFs and IFGSs are sampled from these networks. We describe the steps in Algorithms A.2 and A.3. The two algorithms are for each network. This procedure has been modified from the vanilla Depth First Search Algorithm. To sample IFs and IFGSs, Network2GeneSets (Algorithm A.1) is run on the BFS-Forest F.

**Algorithm A.2** BFS-Forest

**Input:** A $d \times d$ adjacency matrix $A$.
**Output:** A $d \times d$ BFS-Forest adjacency matrix $F$.
Remove all self-transitions in $A$.
Find all of the roots of $A$ and store them in a vector $R$.
**if** no roots **then**
    Sort all vertices in descending order based on their out degree and store them in $R$.
**end if**
Initialize a $d \times d$ Boolean adjacency matrix $F$ with all entries set to **false**.
Initialize a $1 \times d$ vector $nV$ to keep track of the not visited vertices with all elements set to **true**.
**for** each vertex $r \in R$ **do**
    **if** $nV(r)$ is **true then**
        set $nV(r)$ to **false**.
        BFS-Visit(r).
    **end if**
**end for**
Return $F$.

**Algorithm A.3** BFS-Visit

**Input:** A vertex $r$.
**Output:** The updated matrix $F$.
Initialize a queue $Q$ with the vertex $r$ at its head.
**while** $Q$ is not empty **do**
    Pop a vertex $v$ from $Q$.
    Find all of the neighbors $N$ of $v$.
    **if** $N$ is empty **then**
        continue
    **end if**
    **for** each neighbor $n \in N$ **do**
        **if** $nV(n)$ is **true then**
            set $nV(n)$ to **false**.
            Add $n$ to $Q$.
            set $F(v, n)$ to **true**.
        **end if**
    **end for**
**end while**

## A.9 Accommodation of Discrete Inputs by GSGS and GSSA

Let us assume that there are $m$ gene sets and $n$ distinct genes in a given IFGS compendium. Then, the input for GSGS and GSSA can be represented as an $m \times n$ matrix. If there are $k$ genes in the $i^{th}$ gene set, then the corresponding $k$ locations in the $i^{th}$ row contain non-zero indices representing these genes, and the remaining $n - k$ locations are set to zero. Since GSGS and GSSA only considers non-zero indices in a row (or genes present in a gene set), for simplicity we use the first $k$ locations in the $i^{th}$ row to place the non-zero indices and the remaining $n - k$ locations are set to 0. A matrix of this form can be be given as input to GSGS and GSSA.

For Bayesian network and mutual information methods, we construct an $m \times n$ binary data matrix corresponding to the given IFGS compendium. If there are $k$ genes in the $i^{th}$ gene set, then the corresponding $k$ locations in the $i^{th}$ row of data are set to 1 and the remaining $n - k$ locations are set to 0. Depending on the platform, one may have to use $\{1, 2\}$ instead of $\{0, 1\}$ as labels in the binary data matrix. Binary data matrix can be accommodated by the Bayesian network methods K2 and MCMC implemented in the Bayes Net Tool Box (BNT) [101].

Binary data can also be used to infer MI networks. For this, we use the functionalities available in the packages C3NET [8] and MINET [97] available from CRAN/ Bioconductor. For example, the MINET package provides functionalities to estimate mutual information between discrete random variables. The MI estimators for discrete random variables implemented in the package are: Empirical estimator, Miller-Madow corrected estimator, Shrink entropy estimator and the Schurmann-Grassberger estimator. As described in [97], the following steps are involved in the inference of MI network from discrete data:

1. Estimation of mutual information matrix (MIM). Usage: `mutinformation(dataset, estimator)`, where `dataset` is discrete data set and `estimator` is the mutual information estimator.

2. Network inference using ARACNE/C3NET/CLR/MRNET/RN. Usage: `aracne(mim, eps)`, `c3(mim)`, `clr(mim)`, or `mrnet(mim)`. Here `eps` is the threshold used when removing an edge by ARACNE. MIM represents RN.

3. Normalization of the network (optional). Usage: net/max(net).

## A.10 Burn-In State Analysis for GSGS

A burn-in state in Algorithm 5.1 refers to a stage after which we start collecting signaling pathway samples. Samples collected after burn-in state are assumed to be drawn from the joint distribution of IFGSs. To determine an appropriate burn-in state, we translated the approach presented in [40, 41] in our framework to compute the ratio

$$R = \frac{\frac{N-1}{N}W_v + \frac{1}{N}B_v}{W_v} \tag{A.12}$$

for each of the three quantities Sensitivity, Specificity and PPV. Here, $N$ is the total number of pathways sampled after burn-in state, $W_v$ is the averaged within-chain variance (within a single run of GSGS) and $B_v$ is between-chain variance (between multiple runs of GSGS).

Let us fix the burn-in state as $B$ in a total of $J(\geq 2)$ independent runs of GSGS. For a parameter of interest $X$, $W_v$ and $B_v$ are defined as

$$W_v = \frac{1}{J}\sum_{j=1}^{J} s_j^2 \quad \text{and} \quad B_v = \frac{N}{J-1}\sum_{j=1}^{J}(\overline{x}_j - \overline{x})^2,$$

$$\text{where } \overline{x}_j = \frac{1}{N}\sum_{t=B+1}^{B+N} x_j^{(t)},\ s_j^2 = \frac{1}{N-1}\sum_{t=B+1}^{B+N}(x_j^{(t)} - \overline{x}_j)^2,\ j = 1, \ldots, J, \text{ and } \overline{x} = \frac{1}{J}\sum_{j=1}^{J}\overline{x}_j$$

If all the chains are stationary then the numerator and denominator in Eq. A.12 estimate the variance of $X$. It is clear that $\sqrt{R} \to 1$ as $N \to \infty$. In practice, the choice of $B$ and $N$ is acceptable if $\sqrt{R} < 1.2$. Otherwise, either $B$ or $N$ or both should be increased (see [40, 41] for more details).

In our simulation study, we treat Sensitivity, Specificity and PPV as three parameters and compute $R$ in each case. In Case Study I presented in Chapter 5, we considered every $k^{th}$ network among 500 networks sampled after burn-in, for $k = 2, \ldots, 10$. The computations were based on 20 independent runs of GSGS. Under this setting, $\sqrt{R}$ was found approximately equal to one, for each of the three quantities of interest. However, we did not observe a significant change by summarizing sensitivity, specificity and PPV from all 500 networks. It was also observed that there is no much variation in $W_v$ calculated using the networks sampled after burn-in state in different GSGS runs.

## A.11 Computational Complexity Analysis of GSSA

In this section, we first derive the computational complexity of GSSA. We then present numerical results comparing the performance of GSSA and Bayesian network methods in terms of F-score and computational time. Let us write an IFGS compendium as an $m \times n$ matrix, where $m$ is the number of information flow gene sets (IFGSs) and $n$ is the number of distinct genes in the compendium. As described above, if there are $k$ ($k \leq n$) genes in the $i^{th}$ gene set, then the first $k$ locations in the $i^{th}$ row contain non-zero indices representing these genes, and the remaining $n - k$ locations are set to zero. The length of the $i^{th}$ IFGS is the number of non-zero indices in the $i^{th}$ row. If $L$ is the maximum length of IFGSs in the compendium, then the computational complexity of GSSA in taking a total of $J$ jumps is $\mathrm{O}(JmL)$. We can derive the computational complexity of GSSA from Algorithm 6.1 presented in the main text.

We start with the computational complexity involved in calculating the energy of a signaling pathway structure. It is the sum of:

1. The computational complexity of estimating the initial probability vector, which is $O(m)$. This is because we only need to count the frequency of genes appearing as the first node among $m$ Markov chains.

2. The computational complexity of estimating the transition probability matrix, which is $O(mL + n) = O(mL)$. Indeed, we first compute the frequency counts of various transitions among $m$ Markov chains, followed by a normalization of each row in the transition matrix. Moreover, $n \leq mL$.

3. The computational complexity involved in computing the likelihood of a Markov chain, which is $O(L)$. For $m$ chains, the complexity is $O(mL)$.

Thus, the computational complexity of calculating the energy of a signaling pathway structure is $O(mL)$. It can be observed from the pseudo-code in Algorithm 6.1 that the total computational complexity of GSSA depends on the following computations:

### *Outside the loop (Before Step 4)*

1. We need to calculate the lengths of IFGSs and the maximum of the lengths. As we only consider non-zero indices in the given matrix, the worst case computational complexity involved in these computations is $O(mL + m) = O(mL)$.

2. At Step 3, we assign random gene orderings to each of the $m$ gene sets and calculate the energy of the resulting structure. The worst case complexity involved in each of these computations is $O(mL)$.

Thus, the total complexity before Step 4 is $O(mL)$.

### *Inside the loop (Step 4 onwards)*

To jump from $j^{th}$ to $(j+1)^{th}$ network,

1. We need to consider the complexity involved in generating a network from the neighborhood of $j^{th}$ network. Since this requires sampling an index $i \in \{1, \ldots, m\}$ and permuting the order of genes in the $i^{th}$ IFGS, the worst case computational complexity is $O(L)$.

2. We need to consider the complexity involved in calculating the energy of the neighboring network chosen for evaluation, which is $O(mL)$.

Thus, the total computational complexity involved in (1) and (2) above is $O(mL)$, and for a total of $J$ jumps it is $O(JmL)$. As a result, the overall computational complexity outside and inside the loop in Algorithm 1 is $O(mL) + O(JmL) = O(JmL)$.

In Tables A.1-A.4, we present the computational time and performance of GSSA and two Bayesian network methods K2 and MH using IFGS compendiums of different sizes. Unlike MI based algorithms, both GSSA and Bayesian network methods use search strategies for learning multivariate dependencies. Also, both GSSA and Bayesian network methods infer directed network topologies. Therefore, it is relevant to compare GSSA and Bayesian network methods in terms of performance and search time. MI based algorithms are computationally more efficient than GSSA and Bayesian network methods. However, they are suitable for inferring undirected pairwise dependencies among genes.

We use 4 IFGS compendiums among 83 compendiums used in Case Study I. For each algorithm, we list the type of output, computational time and F-Score. As both GSSA and MH depend on the number of jumps/samples specified by the user, we report the performance of these approaches at iteration $10^3, 10^4, 10^5$ and $2 \times 10^5$. We suffix the F-Score ($F$) and elapsed time ($T$) accordingly. Since this is not applicable in the case of K2, we report the final values of $F$ ($F_{Final}$) and $T$ ($T_{Final}$).

We observe from Tables A.1 - A.4 that GSSA benefits from manageable computational complexity and significantly better performance than Bayesian network methods. In particular, GSSA has a much reduced computational load, both in terms of time and memory requirements, compared with sampling based MH algorithm. Note that both GSSA and MH depend on the number of jumps/samples specified by the user. However, the complexity of MH is often unmanageable due a large number of neighboring structures of a sampled network. GSSA only needs to keep track of the best-so-far structure and can be run on a standard desktop.

| Method | Output Type | $F_{10^3}$ | $F_{10^4}$ | $F_{10^5}$ | $F_{2\times10^5}$ or $F_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.57 | 0.89 | 1 | 1 |
| MH-BIC | Directed | 0.21 | 0.27 | 0.45 | 0.49 |
| MH-BAYES | Directed | 0.11 | 0.16 | 0.17 | 0.21 |
| K2-BIC | Directed | * | * | * | 0.41 |
| K2-BAYES | Directed | * | * | * | 0.32 |

| Method | Output Type | $T_{10^3}$ | $T_{10^4}$ | $T_{10^5}$ | $T_{2\times10^5}$ or $T_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.02 | 0.18 | 1.9 | 3.7 |
| MH-BIC | Directed | 0.52 | 5.1 | 51.22 | 103.68 |
| MH-BAYES | Directed | 0.49 | 5.14 | 53.37 | 118.06 |
| K2-BIC | Directed | * | * | * | 0.07 |
| K2-BAYES | Directed | * | * | * | 0.10 |

Table A.1: Comparison of GSSA and the Bayesian network methods in terms of F-Score (Upper Panel) and computational time (Lower Panel). We used IFGS compendium with 54 IFGSs. The lengths of IFGSs varied in the range $4-8$. Time is shown in minutes. Here '*' means Not Applicable.

| Method | Output Type | $F_{10^3}$ | $F_{10^4}$ | $F_{10^5}$ | $F_{2\times10^5}$ or $F_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.69 | 0.91 | 1 | 1 |
| MH-BIC | Directed | 0.09 | 0.22 | 0.30 | 0.34 |
| MH-BAYES | Directed | 0.08 | 0.11 | - | - |
| K2-BIC | Directed | * | * | * | 0.28 |
| K2-BAYES | Directed | * | * | * | 0.20 |

| Method | Output Type | $T_{10^3}$ | $T_{10^4}$ | $T_{10^5}$ | $T_{2\times10^5}$ or $T_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.03 | 0.32 | 3.2 | 6.5 |
| MH-BIC | Directed | 2.6 | 25.15 | 244.15 | 499.59 |
| MH-BAYES | Directed | 2.12 | 27.02 | Out of Memory | Out of Memory |
| K2-BIC | Directed | * | * | * | 0.22 |
| K2-BAYES | Directed | * | * | * | 0.27 |

Table A.2: Comparison of GSSA and the Bayesian network methods in terms of F-Score (Upper Panel) and computational time (Lower Panel). We used IFGS compendium with 108 IFGSs. The lengths of IFGSs varied in the range $4-7$. Time is shown in minutes. Here '*' means Not Applicable and '-' indicates that F-Scores could not be observed due to memory crash.

| Method | Output Type | $F_{10^3}$ | $F_{10^4}$ | $F_{10^5}$ | $F_{2\times10^5}$ or $F_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.45 | 0.54 | 0.63 | 0.74 |
| MH-BIC | Directed | 0.17 | 0.39 | 0.46 | 0.47 |
| MH-BAYES | Directed | 0.09 | 0.14 | - | - |
| K2-BIC | Directed | * | * | * | 0.51 |
| K2-BAYES | Directed | * | * | * | 0.61 |

| Method | Output Type | $T_{10^3}$ | $T_{10^4}$ | $T_{10^5}$ | $T_{2\times10^5}$ or $T_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.04 | 0.39 | 3.9 | 7.9 |
| MH-BIC | Directed | 2.57 | 24.95 | 258.28 | 485.96 |
| MH-BAYES | Directed | 2.22 | 21.11 | Out of Memory | Out of Memory |
| K2-BIC | Directed | * | * | * | 0.26 |
| K2-BAYES | Directed | * | * | * | 0.32 |

Table A.3: Comparison of GSSA and the Bayesian network methods in terms of F-Score (Upper Panel) and computational time (Lower Panel). We used IFGS compendium with 195 IFGSs. The lengths of IFGSs varied in the range $4-10$. Time is shown in minutes. Here '*' means Not Applicable and '-' indicates that F-Scores could not be observed due to memory crash.

| Method | Output Type | $F_{10^3}$ | $F_{10^4}$ | $F_{10^5}$ | $F_{2\times10^5}$ or $F_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.33 | 0.48 | 0.64 | 0.71 |
| MH-BIC | Directed | 0.03 | 0.11 | - | - |
| MH-BAYES | Directed | 0.02 | - | - | - |
| K2-BIC | Directed | * | * | * | 0.30 |
| K2-BAYES | Directed | * | * | * | 0.24 |

| Method | Output Type | $T_{10^3}$ | $T_{10^4}$ | $T_{10^5}$ | $T_{2\times10^5}$ or $T_{Final}$ |
|---|---|---|---|---|---|
| GSSA | Directed | 0.20 | 2.00 | 19.91 | 39.92 |
| MH-BIC | Directed | 380.54 | 2472.71 | Too long | Too long |
| MH-BAYES | Directed | 367.52 | Out of Memory | Out of Memory | Out of Memory |
| K2-BIC | Directed | * | * | * | 11.45 |
| K2-BAYES | Directed | * | * | * | 14.99 |

Table A.4: Comparison of GSSA and the Bayesian network methods in terms of F-Score (Upper Panel) and computational time (Lower Panel). We used an IFGS compendium with 723 IFGSs. The lengths of IFGSs varied in the range $4-12$. Time is shown in minutes. Here '*' means Not Applicable and '-' indicates that F-Scores could not be observed due to memory crash or large computational time.

# Vita

Lipi Rani Acharya received the MSc and PhD degrees in Mathematics from Indian Institute of Technology Madras (2003) and Indian Institute of Technology Kanpur (2009), respectively. In the year 2008, she joined Dr. Zhu's group in the Department of Computer Science at the University of New Orleans for her doctoral study in the field of computational biology. She has been a recipient of the Crescent City doctoral scholarship at the University of New Orleans since 2008. Her research focus is reverse engineering of gene regulatory networks and development of methodologies for pattern discovery from high dimensional molecular profiling data. After the completion of her doctoral study, she will be working as a computational biologist at Dow AgroSciences LLC.